Universidade de Coimbra
Faculdade de Ciências e Tecnologia
Departamento de Engenharia Química

Monitorização, Modelação e Melhoria de Processos Químicos: Abordagens Multiescala Baseadas em Dados

Título em Inglês:

Data-Driven Multiscale Monitoring, Modelling and Improvement of Chemical Processes

Marco Paulo Seabra dos Reis

(Licenciado em Engenharia Química)

Dissertação submetida à Universidade de Coimbra para obtenção do Grau de Doutor em Engenharia Química, na especialidade de Processos Químicos.

Supervisor: Professor Doutor Pedro Manuel Tavares Lopes de Andrade Saraiva

Coimbra, Novembro de 2005

Portugal

Para a Ivone

Abstract

Processes going on in modern chemical processing plants are typically very complex, and this complexity is also present in collected data, which contain the cumulative effect of many underlying phenomena and disturbances, presenting different patterns in the time/frequency domain. Such characteristics motivate the development and application of data-driven multiscale approaches to process analysis, with the ability of selectively analyzing the information contained at different scales, but, even in these cases, there is a number of additional complicating features that can make the analysis not being completely successful. Missing and multirate data structures are two representatives of the difficulties that can be found, to which we can add multiresolution data structures, among others. On the other hand, some additional requisites should be considered when performing such an analysis, in particular the incorporation of all available knowledge about data, namely data uncertainty information.

In this context, this thesis addresses the problem of developing frameworks that are able to perform the required multiscale decomposition analysis while coping with the complex features present in industrial data and, simultaneously, considering measurement uncertainty information. These frameworks are proven to be useful in conducting data analysis in these circumstances, representing conveniently data and the associated uncertainties at the different relevant resolution levels, being also instrumental for selecting the proper scales for conducting data analysis.

In line with efforts described in the last paragraph and to further explore the information processed by such frameworks, the integration of uncertainty information on common single-scale data analysis tasks is also addressed. We propose developments in this regard in the fields of multivariate linear regression, multivariate statistical process control and process optimization.

The second part of this thesis is oriented towards the development of intrinsically multiscale approaches, where two such methodologies are presented in the field of process monitoring, the first aiming to detect changes in the multiscale characteristics of profiles, while the second is focused on analysing patterns evolving in the time domain.

Keywords: Multiscale analysis; Multiscale statistical process control; Measurement uncertainty; Linear regression; Chemometrics; Multivariate statistical process control; Process optimization; Latent variable modelling; Missing data; Multiresolution data.

Preface

“Need is the mother of all inventions”, and the same applies to the underlying motivations of the work presented in this thesis. Several years ago we found ourselves very often is situations where we were confronted with industrial data sets composed of a relatively high number of variables, rather unstructured, noisy and very sparse, with the aim of trying to extract any sort of useful knowledge regarding particular problems that were concerning the process engineers at the moment. By that time there was hardly any tool available ready to be applied (and in fact the situation is not much different now, at least regarding toolboxes commercially available), and, of course, the only way out was to develop and apply alternative approaches tailored to those types of datasets. At the beginning these approaches were rather focused on the specific data sets they were designed to handle, but after some time they evolved to more general data analysis structures, that analyse the information content at different time-scales, looking for the minimum scale where the problem could be tackled, and see what is also contained in the higher scales, using coarser resolution versions of the data sets.

Step by step, the initial approaches gave rise to more structured and general platforms, and some conceptual work began being carried out to provide the necessary theoretical insight on their use: what seemed to be initially a “course” (“why don’t we get the same nice data sets we see in the tutorial books and literature?”…), turned out to be an interesting source of problems still lacking adequate solutions, and, somehow, we now realize what Isaac Asimov meant when he said: “The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' (I found it!) but 'That’s funny …' ”. After some time, the idea of bringing data uncertainty to scene occurred, and it turns out to be instrumental in multiscale platforms where it plays an important role in handling, in a coherent and unified way, the missing data problem and the existence of observations with different qualities.

After this point, we knew we had a problem and a potential way to work it out, and the research work presented in this thesis naturally evolved from them. The multiscale decomposition frameworks were refined, single scale approaches that take advantage of the information generated by the former were developed, and approaches that consider all the scales simultaneously were considered, specially with process monitoring purposes. This thesis is the result of such a work.

Acknowledgements/Agradecimentos

Many people contributed in different ways for the best of my experience as a graduate student, making also the period of time where the work was carried out a better and enriching experience, really worthwhile. To all of them I am very grateful. There are however some people that have accompanied me closer during this journey or whose contributions and advices have particularly marked its course, to whom I would like to address some words of recognition, certainly not enough to make justice to all they have given, but to include them in a work where, in some sense, they were also involved. I will address to them in their native language, if it is *Portuguese*. Otherwise, my words will be in English.

Aos meus pais, Alcides e Natália, que já me apoiam há muito tempo, continuando a fazê-lo com o mesmo empenho e desprendimento de sempre, agradeço todo esse esforço, pois se os meus projectos se vão tornando em realidade, devo-o em primeiro lugar a quem tudo sempre fez para que assim fosse. O mesmo é extensivo à minha irmã Sara, que apesar de entrar "em cena" um pouco mais tarde, cedo me ensinou, com a sua personalidade e independência, que temos muito a aprender com os (irmãos, mas não exclusivamente…) mais novos se assim quisermos.

À Ivone, a quem dedico esta tese, e devo todo um apoio incondicional, muitas vezes em seu prejuízo, ao que acrescento uma coragem realmente inspiradora e compreensão perante quem, como eu, trabalhou na concretização de um projecto, ainda que isso implicasse atrasar os seus. À Sara, minha filha, que com apenas 3 anos me ensina, e muitas vezes sem recurso a palavras, o que é importante não esquecer.

À Universidade de Coimbra pela oportunidade de conduzir o meus estudos de Doutoramento em tão prestigiosa instituição, e ao Departamento de Engenharia Química da sua Faculdade de Ciências e Tecnologia, por me ter proporcionado as melhores condições possíveis para a sua realização. É claro que o apoio de todos os meus colegas, nomeadamente aqueles cujo exemplo, aquando meus Professores na Licenciatura marcou indelevelmente a minha carreira, foi, a todos os níveis, importante para a concretização desta tese. Também interessantes e profícuas foram as inúmeras discussões e (os não tão frequentes, infelizmente, mas mesmo assim excelentes) dias abertos do G3 no velho "Chimico" com o Lino Santos, Eduardo Trincão e Jorge Lemos, e as sessões de almoço às quartas-feiras, com o Fernando Bernardo (a quem

também agradeço a rápida introdução ao formalismo de optimização na presença de incertezas e a companhia amiga nos períodos de descompressão diários), Luísa Durães, Paula Egas e Margarida Quina, bem como os esporádicos "chás das 5", onde também se podia contar com a presença, mais ocasional mas sempre oportuna, do Pedro Nuno Simões e Paulo Jorge Ferreira, onde, em conjunto, revíamos o "estado da nação científica" (e não só). Ao Professor Doutor José Almiro e Castro, por me ter iniciado e apoiado nos primeiros passos da investigação científica. Também gostaria de deixar uma palavra de agradecimento a todos os elementos do GEPSI – PSE group, em particular àqueles com quem interagi mais proximamente, sempre de uma forma enriquecedora para todas as partes, nomeadamente: Raquel Costa, Dina Angélico, Belmiro Duarte, Cristina Gaudêncio e Paulo Quadros.

À Fundação para a Ciência e Tecnologia, pelo apoio financeiro prestado através do projecto POCTI/EQU/47638/2002, fundamental para a concretização dos objectivos da tese.

Ao grupo Portucel Soporcel, e em particular ao Engenheiro José Ataíde, pelos constantes desafios lançados e apoio disponibilizado, à Engenheira Cidália Torre, pelo seu dinamismo e colaboração activa, bem como aos restantes elementos da equipa de Desenvolvimento de Produtos, Engenheira Maria José e Engenheiro Davide Bogas, pelo apoio efectivo sempre que requerido, apesar de todas as demais exigências do dia-a-dia.

Ao Prof. Doutor José Cardoso de Menezes, pela exemplar atitude proactiva no projecto "Hibridar", e aos seus colaboradores, nomeadamente ao João Lopes e ao Vítor Lopes, pelas sempre interessantes trocas de ideias. O mesmo reconhecimento é extensivo aos Prof. Doutores José Sarsfield Cabral e José António Faria, pelos contactos e colaboração mantidos no âmbito do mesmo projecto.

To Professor Bhavik R. Bakshi, for hosting enriching and inspiring visits to its research group at the Chemical Engineering Department of The Ohio State University, and for all the insightful talks we have had, as well as to all the graduate students of the group I had the opportunity to meet during these periods, in particular, Wen-Shiang Chen, Oscar Lara, Nandan Ukidwe, Jorge Hau, and, more recently, Hongshu Chen, for all the warm assistance and fruitful interaction.

To the members of the proENBIS and ENBIS group, for all the learning opportunities and the valuable thoughts they generously shared, in an informal way, regarding their extensive experience on applied statistics, in particular to Tony Greenfield, Andrea Ahlemeyer-Stubbe, Christopher McCollin, the late Professor Dimitar Vandev (who, with a single sentence, inspire the solution to a problem that was concerning me for a long time), Øystein Evandt, Rainer Goeb, Andras Zepleni, Shirley Coleman, Soren Bisgaard, Xavier Tort Martorell and Poul Thyregod.

To Professor Julian Morris and Elaine Martin for hosting my stay at the Chemical Engineering and Advanced Materials Department of the University of Newcastle upon Tyne, to Daniel Castro for introducing me to the CPACT's team and Ahmed Al-Alawi, for all the interesting talks and collaborative work, which, hopefully, will produce results in the near future.

Guardei para o final o agradecimento àquele que prontamente me acolheu quando procurava "novos rumos" para a minha carreira científica e, sem nunca me indicar qual aquele que devia seguir, me colocou na trilha certa. A ele devo as múltiplas oportunidades e experiências que tive durante estes anos de trabalho árduo, dentro e fora do programa de doutoramento, que me permitiram crescer como profissional e pessoa. Por isso, ao Professor Doutor Pedro Saraiva, gostaria de agradecer todo o empenho e amizade que eu tive o privilégio de receber.

Symbols and Abbreviations

Symbols

diag	Operator such that, when applied to a square matrix produces a vector containing its diagonal elements and that, when applied to a vector, produces a square diagonal matrix with the elements of the vector along the main diagonal
J_{dec}	Decomposition depth in the wavelet decomposition
T^2, Q (or *SPE*)	Monitoring statistics for MSPC based on PCA
T_w^2, Q_w	Monitoring statistics for MSPC based on HLV
$u(X)$	Standard uncertainty of X
vec	Operator that vectorizes a matrix or higher order tensors
$\otimes$	Kronecker product operator

Abbreviations

ARL	Average run length
ART	Adaptive resonance theory
ATS	Average time to signal
BLS	Bivariate least squares
HLV	Heteroscedastic latent variable model
IT-net	Input-training neural network
MLMLS	Maximum likelihood multivariate least squares
MLPCA	Maximum likelihood principal components analysis
MLS	Multivariate least squares
MR	Multiresolution
MRD	Multiresolution decomposition
MSPC	Multivariate statistical process control
PCA	Principal components analysis
PCR	Principal components regression
PLS	Partial least squares or projection to latent structures
rMLMLS	“ridge” MLMLS
rMLS	“ridge” MLS
RMSEP	Root mean square error of prediction

RMSEPW	Weighted root mean square error of prediction
SNR	Signal to noise ratio
SPC	Statistical process control
uPLS	Uncertainty-based PLS
USPC	Univariate statistical process control

Table of Contents

List of Figures

List of Tables

Extended Abstract in Portuguese / Resumo Alargado em Português

Apresenta-se nesta secção, de uma forma resumida, o enquadramento, objectivos e contribuições relativas ao trabalho desenvolvido no âmbito desta tese. Na subsecção seguinte, introduz-se o âmbito geral do trabalho aqui apresentado, e apresentam-se as motivações que lhe estão subjacentes, após o que se definem os respectivos objectivos e enumeram as contribuições desenvolvidas na sua persecução. Estas serão descritas com maior detalhe nas subsecções seguintes, onde os principais resultados obtidos serão também brevemente comentados. Finalmente, resumem-se as principais conclusões relativas às contribuições da presente tese e referem-se possíveis linhas para trabalho futuro, numa óptica de continuidade dos esforços de investigação já desenvolvidos.

Introdução

Refere-se de seguida o âmbito geral onde os trabalhos aqui reportados podem ser enquadrados e as principais motivações subjacentes. Os objectivos que nortearam o desenvolvimento das actividades conduzidas, no âmbito desta tese, são também apresentados, e enumeram-se as suas principais contribuições.

Âmbito e Motivação

A natureza dos processos industriais é, actualmente, muito complexa, e o mesmo se aplica, naturalmente, aos dados que deles são recolhidos, que contêm o efeito cumulativo dos vários fenómenos e perturbações que lhes estão subjacentes, os quais possuem diferentes padrões de localização e dispersão no domínio tempo/frequência. Adicionalmente, existe um conjunto de características que é usual encontrar em bases de dados industriais, e que dificultam a sua análise, nomeadamente:

i) Presença de ruído, não raramente com magnitude relativamente elevada (baixo SNR[1]);

ii) Natureza esparsa (proveniente de variáveis com diferentes taxas de aquisição e dados em falha);

iii) Dados com autocorrelação e comportamento não estacionário;

iv) Presença de um elevado número de variáveis com correlações cruzadas (natureza multivariada ou "giga"-variada);

v) Presença de dados com diferentes resoluções (variáveis contendo médias calculadas com base em janelas temporais de diferentes comprimentos);

vi) Padrões distribuídos por várias escalas temporais, com diferentes localizações e dispersões no domínio tempo/frequência (natureza multiescala).

Neste contexto, a extracção de conhecimento útil, para as mais diversas actividades industriais, com vista à melhoria dos processos, está longe se ser uma tarefa trivial, podendo tais dificuldades afectar todos os níveis da hierarquia de tomada de decisão, desde o nível da operação do processo, passando pelo nível da gestão local de uma determinada unidade fabril, e chegando depois até ao níveis de planeamento estratégico. Estes diferentes níveis de tomada de decisão tendem a usar informação com diferentes níveis de resolução. Por exemplo, ao nível da operação do processo, é necessário aceder a informação "composta" na gama dos minutos a horas, enquanto que o Engenheiro responsável pela unidade fabril tipicamente analisa médias horárias ou diárias, a equipa de planeamento da produção se preocupa com valores deslocalizados em horizontes de tempo que vão do dia ao mês, e os elementos do conselho de administração estão essencialmente interessados nas tendências de médias mensais/anuais.

O desenvolvimento de plataformas de projecção que sejam capazes de representar a informação original com diferentes níveis de resolução,[2] de acordo com o fim a que se

[1] Sigla proveniente da língua inglesa, *Signal to Noise Ratio*, significando uma medida da razão entre a magnitude do sinal e a do ruído que o afecta.

[2] O nível de resolução da informação analisada prende-se com o grau detalhe que contém. Se, com base num sinal recolhido com uma dada taxa de amostragem, se calcular um outro, contendo as médias de dois

destina, num ambiente "hostil", em que os dados apresentam as estruturas complexas anteriormente mencionadas, é pois, a nosso ver, não só conveniente e útil mas também oportuno.

Adicionalmente, uma vez escolhida a resolução a que análise vai ser conduzida, é necessário que as ferramentas a empregar estejam não só preparadas para lidar com as características intrínsecas dos dados, mas também que integrem toda a informação útil disponível sobre os mesmos. Em particular, existe hoje uma tendência crescente no sentido de caracterizar os valores recolhidos relativamente à incerteza que têm associada (ISO, 1993; Lira, 2002), tendência esta que tem vindo a ser incentivada pelas organizações internacionais de normalização.[3] Nestas condições, para além da tabela de dados a uma dada resolução ou escala, existe também disponível uma outra, contendo as incertezas associadas a cada valor, a qual deve ser igualmente incorporada na análise.

Finalmente, existem tarefas que, perante a complexidade inerente aos fenómenos industriais, incorporaram, simultaneamente, as várias escalas na sua análise. Estas abordagens, a que designaremos por *multiescala*, têm a capacidade de estudar as diferentes características dos fenómenos distribuídas pelas várias escalas, de uma forma integrada e coerente.

O trabalho realizado no contexto da presente tese, e as abordagens nela propostas, visam precisamente atacar os problemas delineados nos parágrafos anteriores, sendo estas essencialmente baseadas em dados, por oposição às metodologias baseadas em primeiros princípios, que colocam o seu ênfase no conhecimento detalhado dos mecanismos activos nos processos em análise, e na sua transcrição matemática.

---

em dois pontos, perde-se detalhe, e a sua resolução cai neste caso para metade da original; se o horizonte sobre o qual a média é calculada envolver quatro observações consecutivas, a resolução cai para um quarto da original, e assim sucessivamente. A este processo de decaimento da *resolução* corresponde um outro, inverso, de subida na *escala* em que a informação é analisada.

[3] Ver por exemplo a resolução número 21 do "CEN Technical Board", que, em 2003, decidiu dar seguimento às sugestões do grupo de trabalho CEN/BT WG 122 "Uncertainty of measurement", reportadas no documento BT N 6831.

Objectivos

Na sequência do exposto na subsecção anterior, assumiram-se os seguintes objectivos para a presente tese:

i) Desenvolver plataformas de projecção de dados a diferentes níveis de resolução, que sejam capazes de lidar com dados apresentando estruturas complexas (nomeadamente esparsas) e que integrem, de uma forma adequada, toda a informação disponível (dados e a sua incerteza), permitindo nomeadamente, estender o âmbito de aplicação do conceito de análise multiresolução baseada em onduletas[4] a estas novas situações;

ii) Desenvolver metodologias de análise de dados (a uma só resolução, ou escala, i.e., *monoescala*) com particular relevância no âmbito da Engenharia de Sistemas em Processos e Produtos (ESPP), que integrem nas suas formulações a incerteza associada aos dados;

iii) Propor novas metodologias *multiescala* para a monitorização de processos, e desenvolver as existentes de forma a melhorar o seu desempenho, em determinados contextos de aplicação.

Contribuições

As principais contribuições originais desta tese, decorrentes do trabalho desenvolvido na persecução dos objectivos acima delineados, são as seguintes:

i) Criação de três plataformas de *análise multiresolução* (AMR), que integram informação relativa à incerteza dos dados e abordam o potencial problema da existência de dados em falha, no contexto das quais algumas aplicações foram exploradas, incluindo: projecção de dados e respectivas incertezas a

[4] Adopta-se aqui o termo *onduletas* como tradução do inglês, *wavelets*, ou francês, *ondelettes*, uma vez que é aquela que com mais frequência surge em comunicações científicas na língua Portuguesa (Lopes, 2001; Reis, 2000; Soares, 1997), apesar de outras designações poderem também ser usadas, como por exemplo, *ôndulas* (Crato, 1998).

uma dada escala para análise subsequente; selecção da escala de análise; filtragem de sinais afectados por ruído.

ii) Várias metodologias de análise de dados foram analisadas, do ponto de vista do uso que fazem relativamente ao conhecimento das características do ruído que afecta as diversas variáveis, tendo sido desenvolvidas abordagens que integram, explicitamente, informação sobre a incerteza dos dados, nas seguintes áreas:

   a. *Regressão Linear*. Desenvolveram-se modificações a metodologias já existentes, com vista a integrar a incerteza dos dados de uma forma mais completa: métodos MLMLS (*Maximum Likelihood Multivariate Least Squares*), rMLS (*Ridge Multivariate Least Squares*), MLPCR2 (*Maximum Likelihood Principal Components Regression*) e uPLS1–uPLS5 (várias metodologias que integram incerteza dos dados na abordagem PLS, *Partial Least Squares* ou *Projection to Latent Structures*).

   b. *Optimização de processos*. Foram propostas e estudadas comparativamente diversas formulações de optimização de processos, diferindo no nível de incorporação da informação relativa às incertezas que afectam os valores das variáveis medidas.

   c. *Controlo Estatístico Multivariado de Processos*. Propôs-se um modelo estocástico que congrega a incerteza das medições e a variabilidade do processo, e apresentou-se uma metodologia para estimar os seus parâmetros. Este modelo proporciona o suporte probabilístico para implementar sistemas de controlo estatístico multivariado, usando estatísticas de monitorização, as quais também foram desenvolvidas.

iii) Desenvolveram-se duas abordagens multiescala orientadas para a monitorização de processos químicos:

   a. A primeira visa a *monitorização de perfis*, i.e., da relação entre variáveis de entrada (também designadas por descritores, preditores ou regressores) e saída (ou resposta), em que nas variáveis de entrada figuram normalmente descritores da localização espacial ou temporal a

que cada valor da resposta se refere, em particular aqueles que apresentem padrões localizados no domínio da frequência.

b. A segunda aplicação é dirigida ao caso mais convencional, em que se pretendem detectar padrões anormais ao longo do tempo, e consiste em desenvolver uma abordagem, com base no método MSSPC (*Multiscale Statistical Process Control*), capaz de lidar adequadamente com a presença de dados que apresentam diferentes resoluções, no sentido de melhorar o seu desempenho ao nível da definição de regiões onde ocorreram falhas e da rápida detecção do regresso do processo a condições normais de operação.

No contexto da contribuição iii.a), ela foi aplicada a um caso de estudo relacionado com a monitorização da superfície do papel, onde também se analisou com algum detalhe a estrutura multiescala da superfície do papel, usando ferramentas gráficas e séries (ou sucessões) cronológicas, bem como se explorou a informação disponibilizada pelo equipamento de medição adoptado (perfilómetro) relativamente a um conjunto de parâmetros que caracterizam os fenómenos de rugosidade e ondulação. Estes serviram de base ao desenvolvimento de modelos de classificação da qualidade da superfície do papel no tocante àqueles fenómenos, de uma forma quantitativa e estável, recorrendo a espaços de previsão de baixa dimensionalidade efectiva.

As contribuições aqui enumeradas são descritas com mais detalhe nas subsecções seguintes.

Plataformas de Análise Multiresolução Generalizadas

Uma análise multiresolução (AMR) (Mallat, 1989, 1998) decompõe um dado sinal numa versão mais grosseira do mesmo (i.e., de baixa resolução), em conjunto com os sinais de detalhe relativos a todas as escalas inferiores (que se vão perdendo nas aproximações sucessivas a escalas mais elevadas, as quais possuem menor resolução), e é instrumental quando se pretende focar a análise numa escala em particular. No

entanto, a sua aplicação a dados industriais apresenta frequentemente sérias complicações, uma vez que esta é baseada na decomposição de um sinal através da aplicação da transformada de onduleta,[5] a qual pressupõe, por sua vez, a inexistência de dados em falha. Adicionalmente, tal análise não integra explicitamente informação relativa à incerteza dos dados, a qual pode ser relevante para a análise posterior dos resultados da decomposição, que não estarão de facto completos sem a especificação de tal grandeza.

Neste sentido, propõem-se nesta tese abordagens AMR alternativas que, integrando a incertezas dos dados que decompõem, são também capazes de lidar com dados em falha, estendendo assim o âmbito de aplicação da abordagem convencional para situações mais comuns ao nível da estrutura dos dados com origem industrial.

Considera-se aqui *plataforma de análise multiresolução* um algoritmo que proporciona modos de calcular coeficientes de expansão do tipo dos obtidos por aplicação da transformada de onduleta, em diferentes contextos, conforme a seguir se descreve.

Método 1: Ajustar Coeficientes do Filtro de Acordo com a Incerteza dos Dados

A transformada de Haar, talvez a transformada de onduleta mais simples e conhecida, consiste na implementação sucessiva do seguinte procedimento: para cada sinal de aproximação que sucessivamente se vai obtendo, digamos à escala j (começando pelo

[5] Para uma introdução à teoria das onduletas em Português, consultar Reis (2000) e Soares (1997). Relativamente a textos em Inglês, a literatura disponível é hoje bastante extensa, dela figurando desde textos de cariz mais introdutório (Aboufadel & Schlicker, 1999; Burrus *et al.*, 1998; Chan, 1995; Hubbard, 1998; Walker, 1999), tratamentos mais completos do assunto (Mallat, 1998; Strang & Nguyen, 1997) ou que seguem uma linha mais orientada à descrição dos aspectos matemáticos subjacentes (Chui, 1992; Kaiser, 1994; Walter, 1994), até livros mais aplicados (Chau *et al.*, 2004; Cohen & Ryan, 1995; Motard & Joseph, 1994; Percival & Walden, 2000; Starck *et al.*, 1998; Vetterli & Kovačević, 1995) e textos com um maior nível de sofisticação técnica (Daubechies, 1992), sem esquecer ainda alguns artigos de revisão (Alsberg *et al.*, 1997; Rioul & Vetterli, 1991).

sinal original, à escala $j = 0$), calculam-se os novos coeficientes de aproximação para a escala seguinte, $j+1$, através da média sucessiva de blocos não sobrepostos, constituídos por pares de valores anexos, enquanto os respectivos coeficientes de detalhe são obtidos através da diferença entre esta média (ou coeficiente de aproximação) e o elemento de cada bloco:[6]

$$\begin{aligned} a_{\lceil k/2 \rceil}^{j+1} &= C_{\lceil k/2 \rceil} \cdot \left(a_k^j + a_{k+1}^j \right) \\ d_{\lceil k/2 \rceil}^{j+1} &= C_{\lceil k/2 \rceil} \cdot \left(a_{\lceil k/2 \rceil}^j - a_k^j \right) \end{aligned} \tag{1}$$

onde

$$C_{\lceil k/2 \rceil} = \sqrt{2}/2 \tag{2}$$

sendo a_k^j e d_k^j os coeficientes de aproximação e detalhe à escala j, para um índice de translação k, respectivamente, C o coeficiente do filtro envolvido no cálculo dos coeficientes de aproximação e detalhe, e $\lceil x \rceil$ o menor inteiro, *n*, tal que $n \geq x$. Este procedimento confere igual peso a ambos os valores de cada bloco no cálculo da sua média (coeficiente de aproximação). No entanto, se dispusermos de informação relativa à qualidade de cada um destes valores, o processo pode ser modificado, de forma a fazer reflectir, no coeficiente de aproximação, a incerteza associada a cada dado, dando maior peso àquele que possua menor incerteza associada, o que pode ser feito escolhendo valores diferenciados para os coeficientes de cálculo da média (*C*'s), de acordo com um critério apropriado:

$$a_{\lceil k/2 \rceil}^{j+1} = C_{\lceil k/2 \rceil}^{j+1,1} \cdot a_k^j + C_{\lceil k/2 \rceil}^{j+1,2} \cdot a_{k+1}^j \tag{3}$$

---

[6] Os coeficientes de aproximação e detalhe requerem que as médias e diferenças, respectivamente, sejam escalonadas por um factor $1/\sqrt{2}$, de forma a conservar a energia do sinal, após a transformação de onduleta (relação de Parseval, Kreyszig, 1978; Mallat, 1998).

Neste caso, escolheu-se o critério MVUE (*Minimum Variance Unbiased Estimator*) para a estimativa da média (comum), do qual decorre a seguinte fórmula de cálculo para os coeficientes de cálculo da média:

$$C_{\lceil k/2 \rceil}^{j+1,1} = \frac{1 \Big/ u\left(a_k^j\right)^2}{1 \Big/ u\left(a_k^j\right)^2 + 1 \Big/ u\left(a_{k+1}^j\right)^2} \tag{4}$$

$$C_{\lceil k/2 \rceil}^{j+1,2} = 1 - C_{\lceil k/2 \rceil}^{j+1,1} \tag{5}$$

($u(x)$ representa a incerteza associada a x). A correspondente fórmula para os coeficientes de detalhe é a seguinte:

$$d_{\lceil k/2 \rceil}^{j+1} = C_{\lceil k/2 \rceil}^{j+1,1} \cdot \left(a_k^j - a_{\lceil k/2 \rceil}^{j+1}\right) = C_{\lceil k/2 \rceil}^{j+1,2} \cdot \left(a_{\lceil k/2 \rceil}^{j+1} - a_{k+1}^j\right) \tag{6}$$

a qual apresenta uma forte semelhança com a sua congénere correspondente ao caso de Haar. A incerteza deve também ser propagada através das escalas, o que se consegue aplicando a lei de propagação de incertezas à presente situação (ISO, 1993; Lira, 2002):

$$u\left(a_{\lceil k/2 \rceil}^{j+1}\right) = \sqrt{\left(C_{\lceil k/2 \rceil}^{j+1,1}\right)^2 \cdot u\left(a_k^j\right)^2 + \left(C_{\lceil k/2 \rceil}^{j+1,2}\right)^2 \cdot u\left(a_{k+1}^j\right)^2} \tag{7}$$

$$u\left(d_{\lceil k/2 \rceil}^{j+1}\right) = \sqrt{\left(C_{\lceil k/2 \rceil}^{j+1,1}\right)^2 \cdot \left(u\left(a_k^j\right)^2 + u\left(a_{k+1}^j\right)^2 - 2 \cdot C_{\lceil k/2 \rceil}^{j+1,1} \cdot u\left(a_k^j\right)^2\right)} \tag{8}$$

onde se assume que os erros que afectam observações sucessivas são estatisticamente independentes entre si.

Método 2: Usar a Transformada de Haar, Acomodar Dados em Falha e Propagar Incertezas

Nesta metodologia, ao contrário da anterior, os coeficientes do filtro são mantidos constantes, sendo a incerteza dos dados originais propagada para os coeficientes de aproximação e detalhe correspondentes a escalas superiores, segundo a equação (9), e os dados em falha acomodados mediante a aplicação sucessiva do conjunto de regras definido na Tabela 1, durante a fase de decomposição do sinal.

$$u\left(a_{\lceil k/2\rceil}^{j+1}\right)=u\left(d_{\lceil k/2\rceil}^{j+1}\right)=\sqrt{\left(\sqrt{2}/2\right)^2\cdot u\left(a_k^j\right)^2+\left(\sqrt{2}/2\right)^2\cdot u\left(a_{k+1}^j\right)^2} \tag{9}$$

Tabela 1. Plataformas AMR: regras a aplicar na implementação do *Método 2*.

- **Regra 1.** Ausência de dados em falha $\Rightarrow$ usar Haar e calcular incertezas segundo (9)
- **Regra 2.** $\left\{a_k^j\right\}$ está em falha $\Rightarrow \begin{cases} a_{\lceil k/2\rceil}^{j+1}=a_{k+1}^j, u\left(a_{\lceil k/2\rceil}^{j+1}\right)=u\left(a_{k+1}^j\right) \\ d_{\lceil k/2\rceil}^{j+1}=0, \quad u(d_{\lceil k/2\rceil}^{j+1})=0 \end{cases}$
- **Regra 3.** $\left\{a_{k+1}^j\right\}$ está em falha $\Rightarrow \begin{cases} a_{\lceil k/2\rceil}^{j+1}=a_k^j, u\left(a_{\lceil k/2\rceil}^{j+1}\right)=u\left(a_k^j\right) \\ d_{\lceil k/2\rceil}^{j+1}=0, u(d_{\lceil k/2\rceil}^{j+1})=0 \end{cases}$
- **Regra 4.** $\left\{a_k^j, a_{k+1}^j\right\}$ estão em falha $\Rightarrow \begin{cases} a_{\lceil k/2\rceil}^{j+1}=u\left(a_{\lceil k/2\rceil}^{j+1}\right)=\text{"em falha"} \\ d_{\lceil k/2\rceil}^{j+1}=u(d_{\lceil k/2\rceil}^{j+1})=\text{"em falha"} \end{cases}$

Método 3: Usar um Filtro Correspondente a uma Onduleta Ortogonal e Propagar Incertezas

Apesar de uma certa incerteza afectar sempre os dados observados, particularmente quando estes provêm de processos industriais, nem sempre a ausência de dados se coloca como um problema, podendo existir situações onde dispomos de tabelas de valores completas para análise. Nestas condições, é possível usar os filtros

desenvolvidos para as onduletas ortogonais e tirar partido das suas boas propriedades (e.g. compactação de energia de sinais irregulares, descorrelação, suporte compacto, etc.), fruto de um desenho cuidado e teoricamente orientado dos seus coeficientes. Deve-se no entanto complementar este cálculo com a propagação de incertezas para os coeficientes calculados, o que pode ser conduzido, mais uma vez, aplicando a fórmula de propagação de incertezas. No entanto, para a situação em que a incerteza é constante ao longo do tempo e o ruído independente, este cálculo é particularmente simples, uma vez que se pode demonstrar que a incerteza dos coeficientes calculados é igual à dos dados originais (Jansen, 2001; Mallat, 1998).

O *Método 1*, por um lado, e os *Métodos 2* e *3*, por outro, diferem profundamente na forma como incorporam a informação relativa à incerteza dos dados nas suas plataformas AMR. Um estudo mais cuidado revela (Reis & Saraiva, 2005b), como linhas gerais de orientação para o uso destas metodologias,[7] que o *Método 1* dever ser aplicado quando os sinais subjacentes são constantes (ainda que afectados por ruído) ou seccionalmente constantes (até ao nível de decomposição em que o comportamento seccionalmente constante seja quebrado).

Neste mesmo estudo demonstrou-se a utilidade destas plataformas na implementação de estratégias de filtragem baseadas na eliminação selectiva de coeficientes de onduleta mediante o conhecimento das incertezas que afectam os sinais subjacentes, tendo-se verificado que conduzem a melhores resultados do que as suas congéneres mais correntes, em situações onde a incerteza não é homogénea ao longo do sinal (como acontece, por exemplo, quando o ruído é do tipo proporcional).

Outra área onde estas abordagens se revelaram úteis foi no desenvolvimento de ferramentas que assistem o utilizador na selecção da escala para conduzir uma dada tarefa de análise de dados, mediante a indicação da escala mínima, acima da qual é adequado efectuar uma tal análise, bem como no fornecimento dos dados representados à escala seleccionada, em conjunto com as respectivas incertezas que lhes estão associadas. Neste contexto, foram desenvolvidas ferramentas que auxiliam a escolha da

[7] Em que o critério de qualidade adoptado se baseia na capacidade de aproximação do sinal original projectado em cada escala, $j > 0$.

escala tendo por base critérios centrados em: (i) dados em falha; (ii) incertezas dos dados; (iii) ambos os critérios. A sua utilidade foi ilustrada em situações concretas, onde se analisam dados reais, nomeadamente na identificação da escala mínima para análise dos perfis da superfície do papel obtidos por perfilometria (usando um critério baseado em incertezas) e na selecção da escala de análise para um conjunto de dados relativos à qualidade do papel (critério baseado em dados em falha).

Integração de Informação Relativa à Incerteza dos Dados em Metodologias de Regressão

Uma vez seleccionada a escala apropriada para conduzir a análise de dados, nomeadamente usando as ferramentas apresentadas na secção anterior, é altura de conduzir a referida análise, explorando, se possível, toda a informação disponível sobre os dados. Uma importante parte desta informação diz respeito à incerteza que os afecta, a qual define, em última instância, a qualidade de cada valor usado na análise (Kimothi, 2002), devendo por isso ser nela integrada. Na verdade, no seguimento dos esforços desenvolvidos no sentido de especificar a incerteza que afecta valores obtidos experimentalmente, têm surgido outros, que os tornam consequentes em termos da análise que se faz, a qual passa a considerar explicitamente a incerteza dos dados nas suas formulações (Bro *et al.*, 2002; De Castro *et al.*, 2004; Faber & Kowalski, 1997; Galea-Rojas *et al.*, 2003; Martínez *et al.*, 2000; Martínez *et al.*, 2002a; Martínez *et al.*, 2002b; Río *et al.*, 2001; Riu & Rius, 1996; Wentzell *et al.*, 1997a; Wentzell *et al.*, 1997b; Wentzell & Lohnes, 1999).

Nesta tese foram desenvolvidos esforços neste sentido, nomeadamente no que diz respeito à integração da informação relativa à incerteza dos dados no domínio da regressão linear multivariada. Neste contexto, forma analisadas, desenvolvidas e comparadas várias abordagens para a incorporação explícita de incertezas em metodologias clássicas, como OLS (*ordinary least squares*), PLS, PCR (*principal components regression*) e RR (*ridge regression*). Em particular, as seguintes técnicas foram objecto de estudo (indicando-se com "*" aquelas que constituem contributos originais no âmbito desta tese):

1) OLS (Draper & Smith, 1998);
2) MLS (Martínez *et al.*, 2002a; Río *et al.*, 2001);
3) MLMLS* (Reis & Saraiva, 2004b, 2005c);
4) RR (Draper & Smith, 1998; Hastie *et al.*, 2001);
5) rMLS* (Reis & Saraiva, 2004b, 2005c);
6) rMLMLS* (Reis & Saraiva, 2005c);
7) PCR (Jackson, 1991; Martens & Naes, 1989);
8) MLPCR (Wentzell *et al.*, 1997b)
9) MLPCR1 (Martínez *et al.*, 2002a);
10) MLPCR2* (Reis & Saraiva, 2005c);
11) PLS (Geladi & Kowalski, 1986; Haaland & Thomas, 1988; Helland, 1988, 2001b; Höskuldsson, 1996; Jackson, 1991; Martens & Naes, 1989; Wold *et al.*, 2001);
12) uPLS1* (Reis & Saraiva, 2004b, 2005c);
13) uPLS2* (Reis & Saraiva, 2005c);
14) uPLS3* (Reis & Saraiva, 2005c);
15) uPLS4* (Reis & Saraiva, 2005c);
16) uPLS5* (Reis & Saraiva, 2005c).

Relativamente às técnicas acima apresentadas, os métodos 1–6 consistem na resolução dos problemas de optimização indicados na Tabela 2. Os métodos MLPCR1 e MLPCR2 baseiam-se na substituição do método OLS no passo de regressão envolvendo os *scores* provenientes do modelo PCA e a resposta, que não incorpora explicitamente a incerteza dos dados, pelos métodos MLS e MLMLS, respectivamente, que a levam em conta. Relativamente aos métodos alternativos à metodologia algorítmica PLS, uPLS1 e uPLS2 consistem essencialmente no uso dos métodos BLS (*Bivariate Least Squares*, versão univariável do método MLS) e MLMLS, respectivamente, em lugar do método dos mínimos quadrados clássico (OLS), na resolução dos sucessivos problemas de optimização em que o método PLS pode conceptualmente ser subdividido. Por outro

lado, os métodos uPLS3–uPLS5 têm por base diferentes combinações de metodologias de estimação do subespaço predictivo usado em PLS e cálculos dos *scores* neste subespaço:

- *uPLS3* – estima o subespaço predictivo usando uma metodologia baseada em incertezas e calcula os *scores* usando projecções não ortogonais (também baseadas em incertezas);
- *uPLS4* – estima o subespaço predictivo usando a metodologia SIMPLS (de Jong *et al.*, 2001) e calcula os *scores* usando projecções não ortogonais;
- *uPLS5* – estima o subespaço predictivo usando a mesma metodologia usada em uPLS3, e calcula os *scores* usando projecções ortogonais.

Tabela 2. Formulação dos problemas de optimização subjacentes aos métodos OLS, MLS, MLMLS, RR, rMLS e rMLMLS.

OLS	$\hat{b}_{OLS} = \underset{b=\left[b_0 \cdots b_p\right]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y(i) - \hat{y}(i) \right)^2 \right\}$
MLS	$\hat{b}_{MLS} = \underset{b=\left[b_0 \cdots b_p\right]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \frac{\left(y(i) - \hat{y}(i) \right)^2}{s_e^2(i)} \right\}$
MLMLS	$\hat{b}_{MLMLS} = \underset{b=\left[b_0 \cdots b_p\right]^T}{\arg\max} \Lambda(b)$ $\Lambda(b) = -\frac{1}{2} n \ln(2\pi) - \sum_{i=1}^{n} \ln\left(\sigma_{\varepsilon_i}\right) - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{\left(y(i) - \hat{y}(i) \right)^2}{\sigma_{\varepsilon_i}^2} \right)$
RR	$\hat{b}_{RR} = \underset{b=\left[b_0 \cdots b_p\right]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y(i) - \hat{y}(i) \right)^2 + \lambda \sum_{j=1}^{p} b(j)^2 \right\}$
rMLS	$\hat{b}_{rMLS} = \underset{b=\left[b_0 \cdots b_p\right]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \frac{\left(y(i) - \hat{y}(i) \right)^2}{s_e^2(i)} + \lambda \sum_{j=1}^{p} b(j)^2 \right\}$
rMLMLS	$\hat{b}_{rMLMLS} = \underset{b=\left[b_0 \cdots b_p\right]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \ln\left(s_e(i) \right) + \sum_{i=1}^{n} \frac{\left(y(i) - \hat{y}(i) \right)^2}{s_e^2(i)} + \lambda \sum_{j=1}^{p} b(j)^2 \right\}$

Estas metodologias foram alvo de uma análise comparativa, considerando vários cenários relativos à estrutura das variáveis de entrada (níveis de correlação), à natureza

do ruído e tipos de relações entre estas e a variável de saída (Reis & Saraiva, 2004b, 2004c, 2005c), tendo-se verificado que, para dados gerados a partir de modelos com variáveis latentes (Burnham *et al.*, 1999; MacGregor & Kourti, 1998), as metodologias MLPCR conduzem, em geral, a melhores resultados de previsão (Reis & Saraiva, 2004c, 2005c), e, em particular, a metodologia MLPCR2 apresenta o melhor desempenho. Verificou-se também que os resultados obtidos com o método MLMLS são em geral superiores àqueles obtidos com a técnica MLS, e que o método rMLS melhora os resultados obtidos com MLS quando os regressores estão correlacionados, o que indica uma estabilização efectiva do passo de inversão matricial realizado neste método através de uma metodologia análoga à usada em RR. Por outro lado, se os dados provêem de modelos de regressão linear, métodos do tipo PLS, e nomeadamente uPLS1, já apresentam um melhor desempenho global (Reis & Saraiva, 2004b).

Integração da Incerteza dos Dados na Optimização de Processos

Constituindo uma área importante no contexto da Engenharia Química, a optimização de processos químicos foi também analisada do ponto de vista de avaliar o impacto associado à consideração de diversos tipos de incertezas associadas a fluxos de informação *de* e *para* o processo, nos resultados de optimização obtidos.

A avaliação deste impacto foi concretizada através da resolução de três diferentes formulações de optimização, as quais traduzem diferentes níveis de incorporação de informação relativa às fontes de incerteza presentes no processo. Em termos gerais, o problema abordado por estas formulações pode resumir-se através do seguinte enunciado: "*Calcular os valores óptimos a estipular para as variáveis que constituem o vector de entrada (Z) ("óptimos" no sentido de uma dada função objectivo a especificar, ϕ), para uma dada observação do vector das variáveis de carga (L)*".

Colocam-se no entanto algumas situações pertinentes, quando se considera a presença de incertezas neste contexto, e que interessa detalhar:

- As quantidades medidas (i.e., as variáveis de carga, $\tilde{L}$, e as variáveis de saída, $\tilde{Y}$) são afectadas por ruído (que aqui se considera do tipo aditivo), cujas características estatísticas são definidas pela incerteza que lhes está associada:

$$\begin{aligned} \tilde{L} &= L + \varepsilon_L \\ \tilde{Y} &= Y + \varepsilon_Y \end{aligned} \tag{10}$$

onde as quantidades assinaladas com "~" são relativas aos valores observados, enquanto que L e Y se referem aos correspondentes valores "verdadeiros", os quais são, no entanto, desconhecidos (não acessíveis a um observador externo, conforme ilustrado na Figura 1).

- O valor especificado para uma dada variável manipulada, $\tilde{Z}$, (i.e., o seu *set-point*, definido exteriormente), não corresponde exactamente ao valor que de facto irá actuar sobre o processo, devido à presença de um outro tipo de incerteza, que designaremos por "incerteza de actuação" (aqui também considerada do tipo aditivo), e que faz com que a actuação real seja distinta daquela definida externamente.

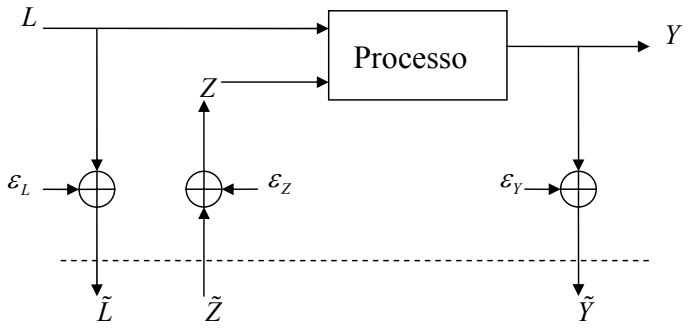


Figura 1. Representação esquemática das quantidades medidas (como são vistas por um operador externo, assinaladas com um "~") e daquelas que de facto interagem com o processo.

Considerando que o objectivo da análise passa pela minimização de uma função custo, $\phi(\cdot)$, propõe-se então a seguinte formulação de optimização, que integra as incertezas associadas aos valores medidos das variáveis de carga e de saída, bem com as incertezas de actuação, a qual consiste na minimização do *valor esperado* da função custo:

Formulação I

$$\begin{aligned} \underset{Z}{Min} \quad & E_{\Theta}\left\{\phi\left(L, Z, \tilde{Y}\right)\right\} \\ s.t. \quad & g\left(Y, L, Z\right) = 0 \\ & L = \tilde{L} - \varepsilon_L \\ & Z = \tilde{Z} + \varepsilon_Z \\ & \tilde{Y} = Y + \varepsilon_Y \end{aligned} \tag{11}$$

onde $E_{\Theta}\{\cdot\}$ é o operador "esperança matemática" e $g\left(Y, L, Z\right) = 0$ representa o modelo do processo, em cujos parâmetros a incerteza se assume desprezável. Nesta formulação, assume-se que o valor medido da variável de saída ($\tilde{Y}$) é uma das quantidades relevantes para o cálculo do custo esperado, o que pode ser justificável nalguns casos, mas deve-se manter presente que podem existir outros onde a quantidade relevante poderá ser porém o próprio valor real, *Y*, (*Formulação 2*), como acontece quando, a jusante, uma medição muito mais rigorosa ficará disponível (proveniente, por exemplo de uma fonte laboratorial ou do fecho de balanços de massa). A formulação correcta a adoptar dependerá por isso da situação particular em causa.

Formulação 2

$$\begin{aligned} \underset{Z}{Min} \quad & E_{\Theta}\left\{\phi\left(L, Z, Y\right)\right\} \\ s.t. \quad & g\left(Y, L, Z\right) = 0 \\ & L = \tilde{L} - \varepsilon_L \\ & Z = \tilde{Z} + \varepsilon_Z \end{aligned} \tag{12}$$

Foi também estudada, para efeitos de comparação com os resultados decorrentes das formulações apresentadas acima, uma terceira formulação, que não considera quaisquer efeitos associados à presença de incertezas, para que melhor se possa avaliar dos potenciais benefícios associados à sua incorporação:

Formulação 3

$$\begin{aligned} \underset{\tilde{Z}}{Min} \quad & \phi\left(\tilde{L},\tilde{Z},\tilde{Y}\right) \\ s.t. \quad & g\left(\tilde{Y},\tilde{L},\tilde{Z}\right)=0 \end{aligned} \tag{13}$$

Estas formulações foram aplicadas a um caso de estudo envolvendo a simulação computacional de um digestor piloto descontínuo de pasta para papel (Carvalho *et al.*, 2003):

$$\text{TY}=55.2-0.39\times\text{EA}+324/\left(\text{EA}\times\log_{10}\text{S}\right)-92.8\times\log_{10}(\text{H})/\left(\text{EA}\times\log_{10}\text{S}\right) \tag{14}$$

onde TY representa o rendimento da pasta (*Total Yield*), EA é o alcali efectivo (*Effectice Alkali*), S é o índice sulfureto e H representa o factor-H usado para o cozimento (para mais detalhes sobre a nomenclatura, consultar Carvalho *et al.*, 2003). A função custo considerada é apresentada na equação (15), a qual penaliza desvios ao valor pretendido para o rendimento ($TY_{sp}=52\%$), levando também em linha de conta os custos associados às variáveis manipuladas S e H. Neste caso de estudo, EA é tomada como sendo a variável de carga, sendo S e H as variáveis manipuladas e TY a variável de saída, i.e., $L=\text{EA}$, $Z=[\text{S H}]$ e $Y=\text{TY}$.

$$L=\begin{cases} 100\left(\dfrac{TY_{sp}}{100}-\dfrac{TY}{100}\right)+\dfrac{S}{4}+\dfrac{H}{500} & \Leftarrow \quad TY\le TY_{sp} \\ 75^2\left(\dfrac{TY_{sp}}{100}-\dfrac{TY}{100}\right)^2+\dfrac{S}{4}+\dfrac{H}{500} & \Leftarrow \quad TY>TY_{sp} \end{cases} \tag{15}$$

Aplicadas à presente situação, as formulações I a III podem ser escritas como indicado na Tabela 3.

Tabela 3. Formulações I, II e III aplicadas à optimização da operação de um digestor descontínuo.

Formulação I	***Formulação II***	***Formulação III***
$\underset{\tilde{S},\tilde{H}}{Min}\ E_\Theta\left\{\phi\left(EA,S,H,\tilde{TY}\right)\right\}$	$\underset{\tilde{S},\tilde{H}}{Min}\ E_\Theta\left\{\phi\left(EA,S,H,TY\right)\right\}$	$\underset{\tilde{S},\tilde{H}}{Min}\ \phi\left(\tilde{EA},\tilde{S},\tilde{H},\tilde{TY}\right)$
$s.t.\quad g\left(TY,EA,S,H\right)=0$	$s.t.\quad g\left(TY,EA,S,H\right)=0$	$s.t.\quad g\left(\tilde{TY},\tilde{EA},\tilde{S},\tilde{H}\right)=0$
$EA=\tilde{EA}-\varepsilon_{EA}$	$EA=\tilde{EA}-\varepsilon_{EA}$	
$S=\tilde{S}+\varepsilon_S$	$S=\tilde{S}+\varepsilon_S$	
$H=\tilde{H}+\varepsilon_H$	$H=\tilde{H}+\varepsilon_H$	
$\tilde{TY}=TY+\varepsilon_{TY}$		

Do estudo decorrente da aplicação destas formulações verificou-se, por exemplo, que o custo associado à consideração dos valores obtidos pela formulação III, quando avaliados à luz da função objectivo para a formulação I, apresenta um agravamento de 136% relativamente àquele obtido pela optimização desta última, sendo o agravamento de 51% quando a formulação II é tomada como referência. A dependência dos resultados perante a magnitude das incertezas que afectam as diferentes variáveis foi também objecto de estudo, tendo-se constatado que, há medida que esta diminui, os resultados (custos) decorrentes das diferentes formulações se aproximam, registando-se adicionalmente uma *diminuição* da função custo, expectável face à diminuição dos efeitos associados aos diversos tipos de ruído que, apesar de serem considerados, impedem cálculos mais precisos das condições óptimas de operação (Reis & Saraiva, 2005c).

Integração da Incerteza dos Dados em Controlo Estatístico Multivariado de Processos

O controlo estatístico multivariado de processos é outra actividade usualmente conduzida a uma só escala (monoescala), onde se procurou integrar o conhecimento relativo à incerteza das medições.

Após uma certa predominância inicial de abordagens univariadas, onde se usavam cartas de controlo desenhadas para seguir o comportamento de variáveis isoladas, constatou-se que tal não constituía uma estratégia eficaz quando aquelas apresentavam dependências ou interacções mútuas, uma vez que tais correlações não eram incorporadas na análise. De facto, a utilização simultânea de várias cartas de controlo univariadas traduz-se numa menor capacidade de detectar acontecimentos especiais, nomeadamente aqueles que violem a estrutura de correlação, sem que os limites de controlo estabelecidos individualmente para cada carta sejam ultrapassados. Para ilustrar esta situação, na Figura 2 encontram-se duas cartas de controlo univariadas, do tipo Shewhart, que monitorizam duas variáveis correlacionadas, X_1 e X_2. Como se pode constatar, nenhuma causa especial de variabilidade é identificada durante o período analisado.

No entanto, procedendo à representação conjunta dos dados para as duas variáveis (Figura 3), facilmente se constata a ocorrência de uma causa especial na décima observação, que passa no entanto completamente despercebida na abordagem univariada. Trata-se de uma observação atípica por violar a estrutura de correlação das variáveis e não devido à variabilidade das suas dispersões marginais. Também se pode verificar que a implementação paralela de cartas de controlo univariadas redunda tacitamente em regiões normais de operação que, no caso multivariável, consistem em hiper-rectângulos, independentemente da forma das funções de densidade de probabilidade.

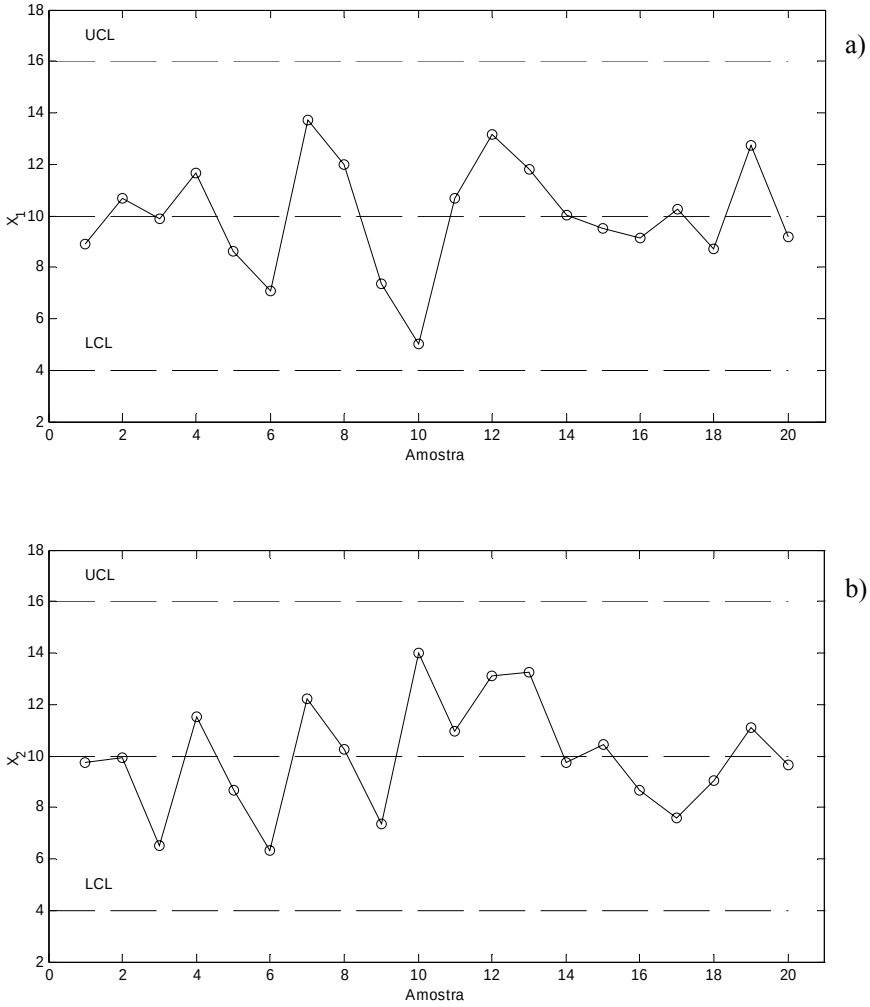


Figura 2. Cartas de controlo univariadas (do tipo Shewhart) para as variáveis X_1 (a) e X_2 (b). UCL (*Upper Control Limit*) e LCL (*Lower Control Limit*) representam limites de controlo, os quais foram especificados para uma região de operação normal correspondente a ± "três–sigma".

Para obviar a estas limitações, desenvolveram-se cartas de controlo multivariadas, que incorporam as correlações existentes entre as variáveis, e dão origem a regiões normais de operação mais adequadas à realidade dos dados, consistindo nomeadamente em hiper-elipsóides, como sucede com a abordagem baseada na estatística T^2 de Hotelling (Montgomery, 2001).

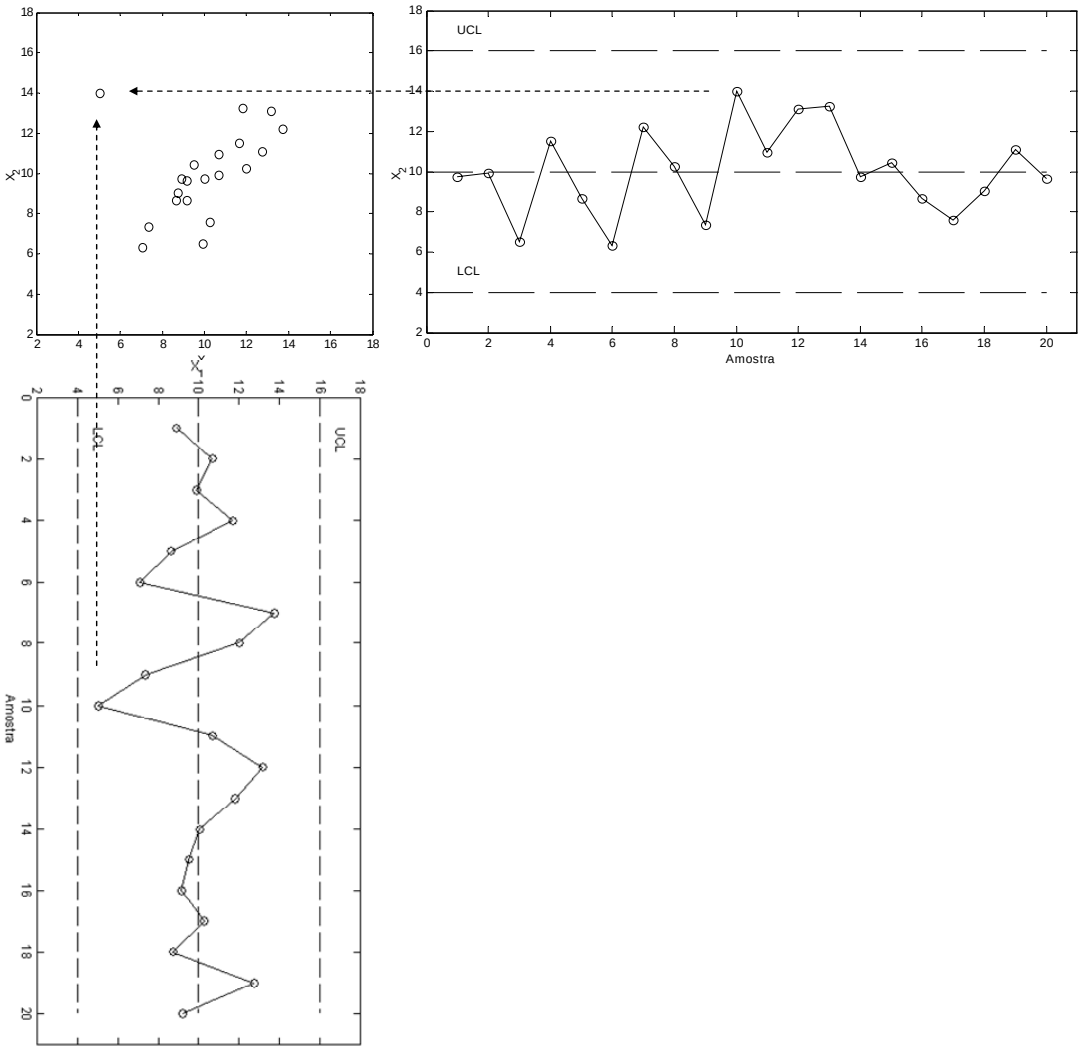


Figura 3. Ocorrência de uma causa especial que não é detectada na abordagem univariada.

No entanto, quando existe um elevado número de variáveis envolvidas, mesmo estas técnicas de controlo estatístico multivariado apresentam problemas, devido ao mau condicionamento da matriz de variância–covariância, a qual deve ser invertida durante a implementação do método (MacGregor & Kourti, 1995). Para contornar esta dificuldade, associada a estruturas de dados mais complexas, surgem as abordagens baseadas em variáveis latentes, especialmente desenhadas para lidar com conjuntos de dados exibindo elevada redundância (Jackson & Mudholkar, 1979; Kourti & MacGregor, 1995; Kresta *et al.*, 1991; MacGregor & Kourti, 1995; Wise & Gallagher, 1996). A implementação destas técnicas baseia-se normalmente em duas estatísticas, através das quais é efectuada a monitorização do estado do processo:

- Uma estatística segue a componente da variabilidade dos dados descrita ou captada pelo modelo de variáveis latentes (e.g. PCA ou PLS), estimado com base em dados históricos relativos a períodos normais de operação (usualmente designada por estatística T^2 ou D);
- A outra estatística segue a variabilidade não explicada pelo modelo, i.e., os resíduos resultantes da projecção dos dados no subespaço correspondente ao modelo de variáveis latentes adoptado (estatística Q ou SPE).

Não obstante a evidente utilidade deste tipo de abordagens em contextos industriais, estas baseiam-se em hipóteses relativamente simplificadas relativamente à estrutura dos erros que afectam as variáveis, as quais se reduzem usualmente ao pressuposto de independência e homogeneidade da variância.

Perante a crescente disponibilidade de informação correspondente à incerteza dos dados, julgamos ser não só oportuno mas também pertinente o desenvolvimento de metodologias de monitorização baseadas em variáveis latentes que incorporem a incerteza dos dados, no sentido de melhorar o seu desempenho em situações onde as variáveis apresentam níveis de incerteza elevados e com variâncias não homogéneas (i.e., não constantes).

Foi neste sentido que se desenvolveu nesta tese uma abordagem para o controlo estatístico multivariado de processos, baseada num modelo de variáveis latentes e que incorpora a incerteza dos dados, de acordo com o seguinte modelo probabilístico:

$$x(k) = \mu_X + A \cdot l(k) + \varepsilon_m(k) \tag{16}$$

onde x é um vector $n \times 1$ contendo as variáveis observadas, μ_X é o vector das médias de x, também $n \times 1$, A é a matriz $n \times p$ dos coeficientes do modelo, l é o vector $p \times 1$ com as variáveis latentes e ε_m o vector $n \times 1$ de erros aditivos, relativos ao ruído de medição (através do qual a incerteza é introduzida nas variáveis observadas). As componentes aleatórias deste modelo seguem as seguintes distribuições:

$$\begin{aligned} &l(k) \sim iid\ N_p\left(\mathbf{0}, \Delta_l\right) \\ &\varepsilon_m(k) \sim id\ N_n\left(\mathbf{0}, \Delta_m(k)\right) \\ &l(k) \text{ e } \varepsilon_m(j) \text{ são independentes } \forall k, j \end{aligned} \tag{17}$$

onde N_p representa a distribuição normal multivariada (p variáveis), Δ_l é a matriz de variância–covariância das variáveis latentes (l), $\Delta_m(k)$ é a matriz de variância–covariância para o vector do ruído das medições no instante k ($\varepsilon_m(k)$), a qual é dada por $\Delta_m(k) = diag(\sigma_m^2(k))$ (sendo $diag(u)$ a matriz diagonal com os elementos do vector u ao longo da diagonal principal e $\sigma_m^2(k)$ o vector contendo as variâncias associadas aos erros no instante k), $\mathbf{0}$ é um vector/matriz de dimensões apropriadas, contendo somente zeros nas suas entradas. Este modelo é designado por HLV (*Heteroscedastic Latent Variable*), sendo constituído por dois blocos fundamentais: um dedicado à descrição da variabilidade normal do processo ($\mu_X + A \cdot l(k)$), e o outro relativo à interferência do ruído de medições ($\varepsilon_m(k)$), cada qual com as suas características aleatórias próprias, aqui tomadas como sendo independentes.

Para implementar uma estratégia de monitorização multivariada com base no modelo HLV (a que designaremos HLV-MSPC, onde MSPC corresponde a *Multivariate Statistical Process Control*), é necessário estimar os parâmetros do modelo (16)–(17), e desenvolver estatísticas adequadas, do tipo das atrás referidas para as estratégias correntes baseadas em variáveis latentes, passíveis de servirem de base a um tal procedimento.

A estimação dos parâmetros é conseguida através da maximização da função log–verosimilhança relevante para o presente caso, i.e.

$$\hat{\theta}_{ML} = \max_{\theta} \Lambda\left(\theta \middle| \left\{x(k), \sigma_m(k)\right\}_{k=1,n_{obs}}\right) \tag{18}$$

com:

$$\theta = \left[\mu_X{}^T, vec\left(\Sigma_l\right)^T\right]^T \tag{19}$$

$$\Sigma_l = A \Delta_l A^T \tag{20}$$

$$\begin{aligned}\Lambda(\mu_X, \Sigma_l) &= -\frac{n \cdot n_{obs}}{2} \ln(2\pi) - \frac{1}{2}\sum_{k=1}^{n_{obs}} \ln\left|\Sigma_x(k)\right| - \frac{1}{2}\sum_{k=1}^{n_{obs}} \left[(x(k) - \mu_X)^T \Sigma_x^{-1}(k) (x(k) - \mu_X) \right] \\ &= C - \frac{1}{2}\sum_{k=1}^{n_{obs}} \ln\left|\Sigma_x(k)\right| - \frac{1}{2}\sum_{k=1}^{n_{obs}} \left[(x(k) - \mu_X)^T \Sigma_x^{-1}(k) (x(k) - \mu_X) \right]\end{aligned} \tag{21}$$

$$\Sigma_x(k) = \Sigma_l + \Delta_m(k) \tag{22}$$

A situação é na verdade um pouco mais complexa, uma vez que a estimativa da matriz Σ_l deve ser simétrica e não-negativa definida (Rao, 1973), o que se consegue através de uma estratégia de optimização na qual, partindo duma estimativa inicial para a matriz de parâmetros, A_0, se procura encontrar a rotação óptima que lhe deve ser aplicada, definida pelo vector de ângulos $\underline{\alpha} = \left[\alpha_1 \ \alpha_2 \cdots \alpha_{n-1}\right]^T$, por forma a maximizar a função (21):

$$\hat{A} = R(\underline{\alpha}) \hat{A}_0 \tag{23}$$

$$R(\underline{\alpha}) = R_1(\alpha_1) \cdot R_2(\alpha_2) \cdot \cdots \cdot R_{n-1}(\alpha_{n-1}) \tag{24}$$

onde

$$R_1(\alpha_1) = \begin{bmatrix} \cos\alpha_1 & -\sin\alpha_1 & 0 & \cdots & 0 \\ \sin\alpha_1 & \cos\alpha_1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \ R_2(\alpha_2) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & \cos\alpha_2 & -\sin\alpha_2 & \cdots & 0 \\ 0 & \sin\alpha_2 & \cos\alpha_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \ etc. \tag{25}$$

Nesta estratégia assume-se que a matriz de variância–covariância para as variáveis latentes é diagonal.

As estatísticas de monitorização, análogas às adoptadas nas metodologias correntes de controlo estatístico multivariado baseadas em variáveis latentes são, na presente situação, as seguintes:

$$\begin{aligned} &T_w^{\,2}(k) = \left(x(k) - \mu_X\right)^T \Sigma_x^{\,-1}(k)\left(x(k) - \mu_X\right) \\ &\Sigma_x(k) = \Sigma_l + \Delta_m(k) \\ &\left(\Sigma_l = A\,\Delta_l\,A^T\right) \end{aligned} \tag{26}$$

e

$$\begin{aligned} &Q_w = r^T(k)\Delta_m^{\,-1}(k)r(k) \\ &r(k) = x(k) - \mu_X - Al(k) = \underline{\varepsilon}_m(k) \end{aligned} \tag{27}$$

onde $T_w^{\,2}(k) \approx \chi^2(n)$ e $Q_w \approx \chi^2(n-p)$, sendo *n* o número de variáveis observadas e *p* o número de variáveis latentes. Os valores de $l(k)$ devem ser calculados com base em "projecções de máxima–verosimilhança" (não ortogonais), usando a seguinte fórmula (Wentzell *et al.*, 1997b):

$$\hat{l}_{ML}(k) = \left(\hat{A}_{ML}^{\,T}\Delta_m^{\,-1}(k)\hat{A}_{ML}\right)^{-1}\hat{A}_{ML}^{\,T}\Delta_m^{\,-1}(k)\left(x(k) - \hat{\mu}_{X,ML}\right) \tag{28}$$

Esta estratégia é estudada recorrendo a vários cenários, envolvendo diferentes estruturas de erros, tendo-se obtido desempenhos de detecção e falsos alarmes, consistentemente superiores ao método convencional nos casos estudados (Reis & Saraiva, 2003, 2005a). Verificou-se também que a incorporação de incertezas na formulação abre uma ponte para a manipulação de dados em falha de uma forma simples, coerente e eficaz, tendo-se registado, na situação analisada, melhores resultados na detecção de situações anómalas relativamente aos obtidos em iguais circunstâncias com a abordagem convencional *sem* dados em falha. O estudo da metodologia HLV-MSPC foi ainda

complementado com a análise de dados reais, provenientes de um processo industrial (Reis & Saraiva, 2005a), aos quais se fez uma análise retrospectiva do histórico disponível, com vista a identificar as principais tendências temporais presentes na variabilidade do processo, permitindo assim adquirir um maior conhecimento sobre o comportamento dinâmico do mesmo em horizontes de tempo mais alargados.

Monitorização Multiescala de Perfis

O problema da monitorização de perfis, i.e., da relação entre variáveis de entrada e saída, em que nas variáveis de entrada figuram normalmente descritores de localização espacial ou temporal (Kang & Albin, 2000; Kim *et al.*, 2003; Woodall *et al.*, 2004), tem vindo a assumir uma importância crescente no domínio do controlo estatístico de processos (Woodall *et al.*, 2004). Neste contexto, desenvolveu-se uma metodologia de monitorização multiescala orientada para este tipo de aplicações, em particular para aqueles perfis que exibem características multiescala, i.e., cuja estrutura apresenta uma dependência da escala onde é analisada, ou, dito de outra forma, cujos fenómenos activos na "construção" do perfil observado apresentam características de localização no domínio da frequência.

Na abordagem proposta, não se considera relevante a descrição de qualquer comportamento localizado no domínio temporal ou espacial, mas somente no domínio que lhes é complementar, da frequência, uma vez que a classe de perfis a que se destina não apresenta tais tipos de padrões, designados por isso de perfis estacionários (no domínio tempo ou espaço, mas não no domínio frequência). Tal abordagem compreende os seguintes passos fundamentais:

1) Aquisição do perfil;
2) Decomposição multiescala do perfil corrigido pela remoção da tendência linear, obtendo-se os coeficientes de onduleta para cada escala ($j = 1 : J_{dec}$, onde J_{dec} é a profundidade da decomposição efectuada);
3) Seleccionar as escalas relevantes para cada fenómeno a monitorar;
4) Utilizando somente as escalas relevantes para cada fenómeno, calcular os parâmetros que sumariam os seus aspectos mais relevantes para efeitos de

controlo de qualidade (este passo pode requerer uma reconstrução separada da parte do perfil correspondente a cada fenómeno no domínio original, através da aplicação da transformada inversa de onduleta a vectores de coeficientes modificados, onde os únicos elementos não nulos correspondem às escalas seleccionadas);

5) Implementar metodologias de controlo estatístico adequadas sobre os parâmetros calculados no passo anterior;

6) Se um alarme for produzido, analisar a sua validade e, se necessário, identificar a suas causas. Caso contrário, regressar ao passo 1 e repetir o procedimento para o próximo perfil adquirido.

Esta abordagem foi testada no âmbito de um caso de estudo que envolveu a caracterização, modelação e monitorização da superfície do papel, no decorrer do qual:

- os fenómenos de rugosidade e ondulação, associados à superfície do papel, foram analisados com base em perfis obtidos por perfilometria e no subsequente estudo, recorrendo, por exemplo, à teoria das séries cronológicas;
- os parâmetros que caracterizam estes fenómenos, fornecidos directamente pelo aparelho utilizado, serviram de base ao desenvolvimento de modelos de classificação que prevêem a classe de qualidade associada a uma determinada folha de papel (no tocante a cada um destes fenómenos), usando como referência classificações previamente efectuadas por um painel de especialistas;
- os perfis captados foram usados para testar o procedimento de monitorização multiescala proposto, em conjunto com outros gerados computacionalmente a partir de modelos realistas da superfície do papel.

Como principais resultados, salienta-se a capacidade de desenvolver modelos de classificação da qualidade superficial do papel com base num espaço de previsão de baixa dimensionalidade (Reis & Saraiva, 2005f) e o bom desempenho da abordagem proposta na monitorização dos fenómenos de rugosidade e ondulação (Reis & Saraiva, 2005d, 2005e).

Controlo Estatístico Multivariado e Multiescala Usando Variáveis com Diferentes Resoluções

A estratégia de controlo estatístico multivariado e multiescala usualmente conhecida por MSSPC (*Multiscale Statistical Process Control*), proposta por Bakshi (1998), mostrou-se adequada para lidar com uma ampla variedade de perturbações que podem afectar os processos, com diferentes características de localização e dispersão nos domínios do tempo e frequência. No entanto, apesar de ser intrinsecamente multiescala, pois analisa separada e simultaneamente a informação distribuída pelas diferentes escalas, esta estratégia assume que toda a informação é disponibilizada a uma só resolução (a mais fina). No entanto, não raramente diferentes variáveis reportam valores correspondentes a médias calculadas em diferentes horizontes do tempo, ou relativos a quantidades de produto/matéria-prima recolhidas ao longo de períodos de tempo, após o que são misturados e analisados. Estas acções geram valores tabelados cuja localização temporal efectiva (resolução) é distinta (estruturas multiresolução, MR), o que levanta dificuldades no seu processamento usando técnicas convencionais, baseadas no pressuposto de resolução única.

Assim, propõe-se nesta tese uma abordagem de controlo estatístico multivariado e multiescala que processa adequadamente dados multiresolução, designada por MR–MSSPC, a qual permite melhorar, relativamente à abordagem convencional, a definição dos períodos de tempo onde efectivamente se localiza uma anomalia, bem como detectar rapidamente o regresso do processo ao estado normal de operação.

Apesar da abordagem proposta se basear numa implementação da transformada de onduleta *ortogonal* numa janela diádica de comprimento variável, a qual introduz algum atraso na disponibilização de coeficientes de onduleta, o desempenho em termos das métricas associadas à rapidez de detecção não é normalmente afectado, a menos que a falha afecte, única e exclusivamente, a(s) variável(is) de baixa resolução, sendo aliás, nas restantes situações, comparativamente melhor, excepto para grandes perturbações, onde pode demorar mais um instante de tempo (no máximo) a detectar a perturbação.

A abordagem proposta foi também aplicada à monitorização de um processo simulado, constituído por um CSTR a operar em regime dinâmico sob controlo retroactivo de temperatura e nível, onde se ilustram vários aspectos pertinentes na sua implementação prática, nomeadamente no que se refere à: selecção da profundidade de decomposição a

usar na transformada de onduleta, e do número de componentes principais a considerar nos modelos correspondentes às diferentes escalas.

Conclusões

A complexidade dos processos industriais, e dos dados deles recolhidos, requer, cada vez mais, plataformas de cálculo adequadas e ferramentas flexíveis na sua análise. Nesta tese, analisaram-se e propuseram-se desenvolvimentos neste contexto, visando a criação de uma abordagem estruturada de análise da informação contida nos dados industriais em várias escalas, capaz de lidar com a presença de dados em falha (uma característica intrínseca dos processos industriais) e de integrar informação relativa à incerteza dos dados recolhidos. Tais plataformas, ditas plataformas AMR generalizadas, permitem, entre outras aplicações, representar a informação a uma escala seleccionada (propagando a incerteza dos dados para a escala em causa) e auxiliar o utilizador na selecção da escala de análise, por sugestão da escala mais fina onde esta pode ser conduzida.

Uma vez seleccionada a escala de interesse, qualquer análise monoescala deve ser conduzida de forma a incorporar toda a informação disponível sobre os dados, nomeadamente a incerteza que lhes está associada. Os estudos conduzidos nesta tese demonstram a pertinência e os ganhos associados à consideração deste importante elemento adicional, nomeadamente nas áreas de regressão linear, optimização de processos e controlo estatístico de processos.

Para as situações em que a complexidade das estruturas de dados ou processos envolvidos requer a análise de várias escala simultaneamente, propuseram-se desenvolvimentos no domínio das abordagens multiescala, os quais permitem: (i) conduzir a monitorização de perfis com características localizadas no domínio da frequência de uma forma eficaz, (ii) integrar informação com diferentes níveis de resolução no método MSSPC, melhorando o seu desempenho na definição de regiões em que ocorrem anomalias e na detecção rápida do regresso ao estado de operação normal.

Trabalho Futuro

A área da análise multiescala de dados e processos apresenta-se em franco desenvolvimento e tem ganho momento em vários domínios do conhecimento. Existem por isso inúmeras ramificações futuras para os esforços desenvolvidos no âmbito desta tese, enunciando-se aqui algumas notas sobre linhas de investigação pertinentes de continuidade ou que representam novos desafios interessantes a considerar:

- *Modelação empírica multiescala*. É amplamente reconhecido o papel fundamental que os "modelos" assumem na Engenharia Química, os quais marcam presença, de uma forma mais ou menos explícita, em praticamente todas as tarefas conduzidas (e.g. controlo, optimização, projecto). Existem situações em que não existe conhecimento suficiente sobre os processos ou fenómenos em curso para que um modelo desenvolvido com base em primeiros princípios produza previsões com uma suficiente aderência à realidade. Nestes casos, os modelos empíricos, baseados em dados, ou os modelos híbridos, baseados na combinação do conhecimento existente com dados recolhidos, constituem abordagens alternativas a considerar. No tocante aos modelos empíricos, as estratégias convencionais são intrinsecamente monoescala (modelos de séries cronológicas, espaço de estados, variáveis latentes), não possuindo por isso a flexibilidade de modelar explicitamente os fenómenos distribuídos pelas várias escalas ou bandas de frequência. Afigura-se pois como pertinente o desenvolvimento de novas estruturas de modelação e métodos de estimação, mais adequadas na descrição de fenómenos multiescala, as quais, uma vez disponíveis, servirão de suporte ao desenvolvimento de novas versões multiescala, congéneres das correspondentes técnicas monoescala (estimação óptima, monitorização, controlo).

- *Monitorização multiescala*. Existem ainda outras metodologias de monitorização multiescala, bem como domínios de aplicação que interessa explorar no futuro. Um exemplo concreto consiste na quantificação da energia localizada nas diferentes escalas, para todas as variáveis a monitorar, e no seu seguimento ao longo do tempo, usando janelas de dados não sobrepostas. A sua distribuição conjunta, em condições normais de operação, permite estabelecer os limites de controlo para as estatísticas de monitorização. Tal abordagem pode ser

aplicada em áreas como a monitorização de processos multivariados ou a detecção de falhas em sensores isolados.

- *Modelação de redes estruturadas multiresolução*. Os Engenheiros Químicos estão, em geral, bem familiarizados com a hierarquia das várias actividades de tomada de decisão numa organização ligada ao sector produtivo. Esta é normalmente representada por uma pirâmide: na base estão situadas as decisões operacionais, relacionadas com a normal laboração dos vários processos industriais; nos níveis intermédios encontra-se aquelas que se prendem com o desempenho da unidade fabril como um todo, planeamento da produção, gestão de disponibilidades de produtos e matérias primas; nos níveis superiores temos as decisões estratégicas, onde essencialmente se estabelecem as directrizes futuras para a organização (planos de investimento e expansão). É interessante notar que, paralelamente a esta pirâmide de tomada de decisão, existe uma outra relativa à resolução típica da informação processada nas decisões tomadas a cada nível: na base da pirâmide, a informação processada é usualmente mais fina (minutos/horas); nos níveis intermédios, trabalha-se com base em médias horárias/diárias (supervisão da unidade fabril) ou diárias/mensais (planeamento de produção); no nível superior tipicamente são analisadas as tendências de indicadores compostos, calculados na base mensal/anual. Constata-se portanto, a existência de uma outra estrutura piramidal, na qual a informação é encaminhada para os níveis superiores num formato sucessivamente mais compacto (de menor resolução), circulando as decisões em sentido oposto, do topo da pirâmide para a sua base. Seria pois interessante, no futuro, traduzir estes elementos numa abordagem de modelação estruturada, abrangendo todas as partes envolvidas na tomada de decisão e incorporando a resolução em que estas de facto operam, por forma a descrever de uma forma mais realista o comportamento global das organizações industriais, utilizando tal conhecimento no estudo de políticas fundamentadas, envolvendo um ou vários níveis da hierarquia, incluindo estudos das cadeias de produção/distribuição, impactes ambientais e sociais, cada vez mais relevantes nos tempos que correm.

- Os esforços desenvolvidos nesta tese, no sentido de integrar a informação relativa à incerteza das medições em diversas áreas de análise de dados, podem e devem ser continuados. Como exemplos de algumas áreas onde tal pode ser

efectuado, referem-se a regressão não-paramétrica, classificação (abordagens paramétricas e não-paramétricas), controlo e estimação.

Part I

Introduction and Goals

A journey of a thousand miles begins with a single step

Lao-tzu (6th century BC, China)

Chapter 1. Introduction

In this introductory chapter, a general perspective of the contents and matters treated in this thesis is provided. It is divided in four separate sections, where its main elements and structure are briefly described. In the first section, motivation to the work carried out is addressed, as well as the scientific scope where the thesis contributions may be considered to belong. Then, in the second section, the main goals are defined and, in the third, the thesis contributions are presented. Finally, in the fourth section, an overview of the thesis' structure is provided.

1.1 Scope and Motivation

Processes going on in modern chemical processing plants are typically complex, and this complexity is also present in collected data, which contain the cumulative effect of many underlying phenomena and disturbances, with different location and localization patterns in the time/frequency plane, as well as a number of additional complicating features that often hinders the analysis of collected data using conventional tools. In particular, industrial data bases typically present the following characteristics:

i) Presence of noise, quite often with low signal to noise ratios (SNR);

ii) Sparse structure (variables with different acquisition rates and with randomly missing blocks);

iii) Multivariate or "giga"-variate with cross-correlations;

iv) Autocorrelation and non-stationary behaviour;

v) Multiresolution data (variables containing averages computed over different time supports);

vi) Multiscale features (phenomena with different patterns in the time/frequency plan);

vii) Variables are not naturally aligned in time;

viii) Presence of corrupted data.

In this context, the extraction of useful knowledge from industrial data has become an increasing complex task, and engineers frequently find themselves in a situation of being "data-miners" (Wang, 1999) that seek knowledge hidden in an immense (large), dirty (corrupted with noise, flaws, irrelevant information, etc.) and hard to handle "data-mine". The above issues can be relevant at all different levels of decision-making: from the *operation* level, where one aims to run the process as smoothly as possible following the operation schedule and the focus is on detecting promptly process faults and special events, passing by the *processing unit management* where the monitoring and control of the overall unit performance take place, and concerns with issues such as reducing product's quality variation arises, moving all the way up to the *strategic* and *planning* levels, that plan the forthcoming production schedule and define general plant policies using data to support their decisions that in this situation also come from other sources, other than plant facilities. At all of these layers, information should be made available in the adequate format, which typically involves the use of different resolutions, reflecting the distinct levels of detail relevant for the analysis undertaken and subsequent decision-making.

Thus, the development of frameworks that are able not only to handle complex features but also to represent data conveniently at the different relevant resolution levels, in an integrated and consistent way, is highly desirable. The possibility of performing a scale-dependent analysis can also help in the identification of those scales where most of the hidden information or critical relationships are established, and should be complemented with adequate single-scale tools, that take the most of the information made available at a particular (selected) scale.

The approaches proposed and analysed in this thesis are directed towards the points raised above, falling under the broad class of data-driven methodologies (Saraiva, 1993; Saraiva & Stephanopoulos, 1998), as opposed to first principles-based methodologies

that rely extensively on the availability of knowledge regarding phenomena going on at different scales and how they interact with each other (Braatz *et al.*, 2004; Charpentier & McKenna, 2004; Li & Kwauk, 2003; Li & Christofides, 2005).

1.2 Goals

In the sequel of the motivating considerations presented in the previous section, the following general tasks/problems were assumed as targets to be accomplished in this thesis:

i) Develop frameworks that are able to perform multiscale or multiresolution decompositions in data structures with complicating features, namely in the presence of missing data and taking into consideration their noise structure, in order to support subsequent data analysis tasks involving several scales (multiscale analysis) or just a particular scale (single-scale[8] analysis).

ii) Develop data analysis tools that are able to take advantage of the type of information generated by the above mentioned frameworks, namely values and associated uncertainties at a given scale.

iii) Propose new and/or improve existent multiscale approaches for process monitoring.

1.3 Contributions

We consider the following as being the main contributions associated with this thesis, relative to the research goals defined previously:

i) Three uncertainty-based multiresolution decomposition methodologies (MRD) are proposed: *Methods 1* and *2* handle the presence of missing data

[8] We will refer in this thesis as "single-scale", those approaches that only deal with data at a single resolution, i.e., without considering either as inputs or in the core of their algorithms any analysis of data segregated according to scale, resolution or frequency.

and any structure of data uncertainties, the former being especially devoted to piecewise constant signals; *Method 3* handles those cases where no missing data is present, incorporating data uncertainty in the computation of detail and approximation coefficients. The problems of scale selection and de-noising are also addressed from the perspective of using the information generated by these frameworks.

ii) Regarding uncertainty-based data analysis tools, the following single-scale methodologies are proposed and compared with others already developed:

   a. *Linear regression*: Maximum Likelihood Multivariate Least Squares (MLMLS), ridge Multivariate Least Squares (rMLS), ridge Maximum Likelihood Multivariate Least Squares (rMLMLS), a modification of maximum likelihood principal components regression (MLPCR2), and five uncertainty-based modifications of partial least squares[9] (uPLS1, uPLS2, uPLS3, uPLS4, uPLS5);

   b. *Process optimization*: several possible optimization formulations were proposed and analysed, differing on the levels of incorporation of uncertainty information;

   c. *Multivariate statistical process control*: a statistical model was proposed (the heteroscedastic latent variable model, HLV) that provides the probabilistic backbone for integrating uncertainty information in multivariate statistical process control (MSPC), as well as an algorithm for estimating its parameters. Associated monitoring statistics were also put forward;

iii) Two multiscale process monitoring approaches were developed

   a. The first methodology regards the multiscale monitoring of profiles, and is built around a wavelet-based multiscale decomposition framework that essentially conducts a multiscale filtering of the raw profile, effectively separating the relevant phenomena under analysis located at different

[9] Also referred to as projection to latent structures (PLS).

scales, allowing also for the incorporation of available engineering knowledge;

b. The second methodology provides a way of conducting MSSPC by adequately integrating data with different resolutions (multiresolution data), in order to improve the definition of the regions where significant events occur under these circumstances, and leads to a more sensitive response when the process is brought back to normal operation.

In the context of contribution iii.a), and, in particular, the case study analyzed to illustrate its application, the monitoring of paper surface, the multiscale structure of paper surface was also carefully analyzed using specialized plots and time series theory, and parameters provided by the measurement device adopted were also used for predicting paper surface quality regarding waviness and smoothness, by developing classification models that adequately explain assessments made by a panel of experts.

In the Future Work chapter, some developments are also proposed and preliminary results presented that demonstrate their potential usefulness, namely regarding another multiscale approach for process monitoring and the integration of uncertainty information in non-parametric regression tasks, which are not included in the body of the thesis, as further testing and analysis must be devoted to them, in order to establish their properties more thoroughly.

1.4 Thesis overview

The present thesis is divided into five distinct parts (Figure 1.1).

Part I provides the necessary motivation and defines the general scope of the work reported in the thesis, as well as establishes the goals and summarizes the main contributions achieved in their persecution.

In Part II a state of the art review regarding multiscale approaches in Chemical Engineering (and closely related fields) is presented.

Part III contains background material necessary to follow the methodologies proposed here, covering subjects like statistical process control, latent variable models, measurement uncertainty, and, in particular, wavelet theory, all of them playing an important role on several parts of the thesis.

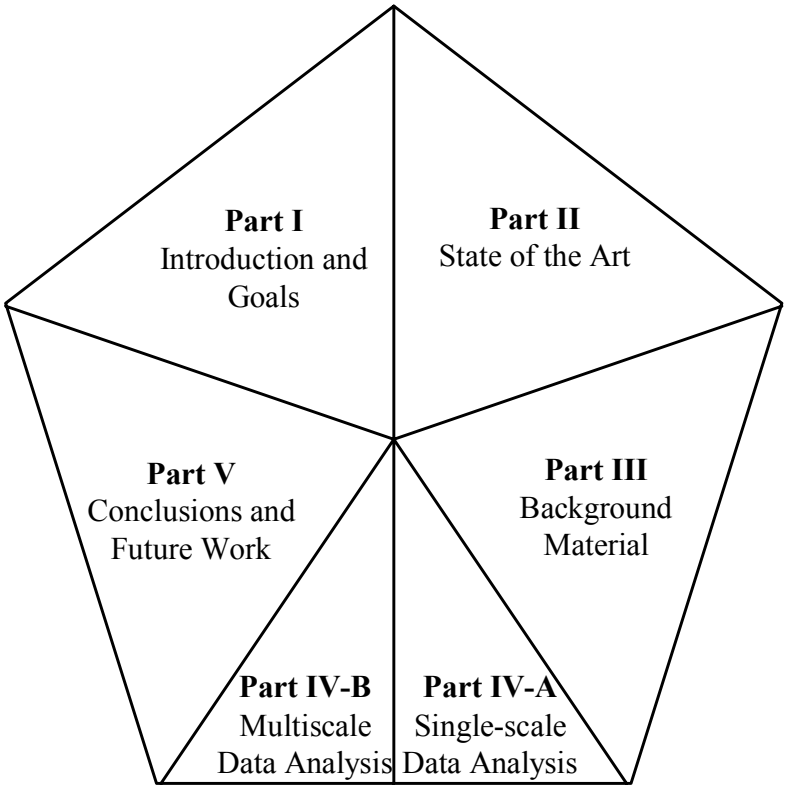


Figure 1.1. The five different parts that compose the thesis.

The original thesis contributions are presented in Part IV-A and Part IV-B. These contributions are divided in two parts, the first (A) devoted to single-scale methodologies, while the second (B) regards inherently multiscale approaches. The first chapter of Part IV-A (Chapter 4) addresses the development of uncertainty-based multiresolution decomposition frameworks that can serve the purposes of both single-scale or multiscale approaches. However, they were presented in this first part, as for some single-scale applications one may find adequate the previous use of one of such frameworks. The following chapters (Chapters 5 to 7) address developments regarding the incorporation of data uncertainty information into several single-scale data analysis tasks, such as: linear regression modelling (Chapter 5), process optimization (Chapter 6) and multivariate statistical process control (Chapter 7).

Part IV-B presents a multiscale approach for monitoring profiles (Chapter 8) and a modification of multiscale statistical process control (Bakshi, 1998), in order to allow for its extension to the situation where multiresolution data is available (Chapter 9).

Finally, in Part V the main conclusions of the present thesis are summarized (Chapter 10) and future work is addressed (Chapter 11).

Part II

State of the Art

If I have seen further it is by standing on ye shoulders of Giants.

Sir Isaac Newton (1642-1727), English mathematician and physicist.

I can't see any farther. Giants are standing on my shoulders!

Unknown

Chapter 2. Applications of Multiscale Approaches in Chemical Engineering and Related Fields: a Review

The multiscale features generated by complex phenomena going on in chemical processing plants, and also present in collected data, call for adequate approaches with the potential for extracting and analysing in effective ways, the information content distributed across the relevant scales. In this context, wavelet theory provides a rich source of tools for supporting multiscale data analysis tasks, when there is a certain lack of fundamental knowledge regarding the underlying phenomena (a situation often found in industrial applications) required to implement first principles-based multiscale modelling and analysis approaches. Therefore, a review of relevant publications in the field of data-driven multiscale analysis is necessarily almost coincident with the one regarding the application of wavelets in chemical engineering, as these are, almost invariably, the workhorse for any of such analysis.

In the following sections of this chapter, a review of relevant publications regarding data-driven multiscale approaches in several research areas from Chemical Engineering and related fields is presented. The approaches to be presented essentially explore some of the properties that make wavelets transforms particularly useful for several data processing and analysis tasks, namely:

1. Wavelet transforms can easily detect and efficiently describe localized features in the time/frequency plane, being by these reasons promising tools for analysing data arising from non-stationary processes or that exhibit localized regularity patterns;
2. They are able to extract the deterministic features in a few wavelet coefficients (energy compaction). On the other hand, stochastic processes spread their energy through all the coefficients and are approximately decorrelated, i.e., the autocorrelation matrices of such signals are approximately diagonalized (decorrelation ability) (Bakshi, 1999; Dijkerman & Mazumdar, 1994; Vetterli & Kovačević, 1995);
3. Wavelet theory provides a framework for analysing signals at different resolutions (the multiresolution decomposition analysis), with different levels of detail (Mallat, 1989);
4. Wavelets provide an efficient representation of both smooth functions and singularities (Burrus *et al.*, 1998);
5. Computations involved are inexpensive (complexity of $O(N)$).

2.1 Signal and Image De-Noising

In spite of not belonging to what is traditionally considered its core, the applications referred in this section and in the next one found many applications in Chemical Engineering, and are furthermore relative to areas where wavelets have become quite popular, owing to the achieved results and simplicity of the methodologies involved. These applications also allow one to develop an intuitive understanding about the reasons why they work so well in some situations, and how we can take advantage of that by extending the successful methodologies to other applications.

In general terms, de-noising concerns uncovering the true signal from noisy data where it is immersed,[10] and is one of the classical application fields where wavelets have found

[10] A more formal interpretation of the term "de-noising" is provided elsewhere (Donoho, 1995).

wide application. The success arises mainly from their ability to concentrate deterministic features in a few high magnitude coefficients while the energy associated with stochastic phenomena is spread over all coefficients. This property is instrumental for the implementation of thresholding strategies in the wavelet domain. Donoho and Johnstone pioneered this field and proposed a simple and effective de-noising scheme for estimating a signal with additive i.i.d. zero-mean Gaussian white noise (Donoho & Johnstone, 1992):

$$y_i = x_i + \sigma \cdot \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0,1) \quad (i = 1,\ldots,N) \tag{2.1}$$

consisting of the following three steps:

1. Compute the Wavelet Transform of the sequence $\{y_i\}_{i=1:n}$ (the boundary corrected or interval adapted filters developed by Cohen, Daubechies, Jawerth and Vial were suggested by the authors);
2. Apply a thresholding policy to the detail wavelet coefficients (the authors suggested soft-thresholding), with $T = \hat{\sigma}\sqrt{2\ln(N)}$, where $\hat{\sigma}$ is an estimator of the noise standard deviation – usually a robust approach is applied to the wavelet coefficients at the finest scale, such as $\hat{\sigma} = Med\left(\left|\left\{d_k^1\right\}_{k=1,\ldots,N/2}\right|\right) \Big/ 0.6745$:

$$\textbf{“Hard-Thresholding”: } HT(x) = \begin{cases} x \Leftarrow |x| > T \\ 0 \Leftarrow |x| \le T \end{cases} \tag{2.2}$$

$$\textbf{“Soft-Thresholding”: } ST(x) = \begin{cases} sign(x) \cdot (|x| - T) \Leftarrow |x| > T \\ 0 \Leftarrow |x| \le T \end{cases} \tag{2.3}$$

3. Compute the inverse Wavelet Transform, thus obtaining the de-noised signal.

This simple scheme is called “VisuShrink”, since it provides better visual quality than other procedures based on mean-squared-error alone. As an illustration, Figure 2.1 depicts the de-noising of an NMR spectrum using a Symmlet-8 filter, for a decomposition depth of $j = 5$.

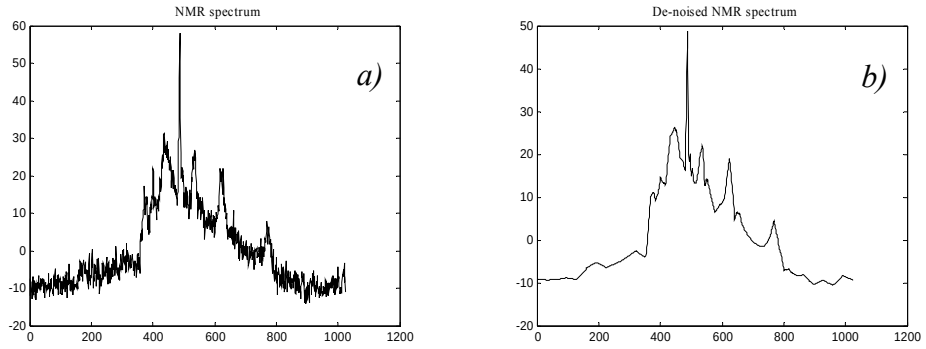


Figure 2.1. De-noising of an NMR spectrum: a) original NMR spectrum; b) de-noised NMR spectrum (WaveLab package, version 8.02, was used in the computations, carried out in the Matlab environment, from MathWorks, Inc.).

This task constitutes in fact a *non-linear* estimation procedure, since the wavelet thresholding scheme is both adaptive and signal dependent, in opposition to what happens, for instance, with optimal linear Wiener filtering (Mallat, 1998) or to thresholding policies that tacitly eliminate all the high frequency coefficients, sometimes also referred to as smoothing techniques (Chau *et al.*, 2004; Depczynsky *et al.*, 1999).

Since the first results published by Donoho and Johnstone, there have been numerous contributions regarding modifications and extensions of the above procedure, in order to improve its performance for a variety of application scenarios. Orthogonal wavelet transforms lack the translation-invariant property and this often causes the appearance of artefacts (also known as pseudo-Gibbs phenomena) in the neighbourhood of discontinuities. Coifman and Donoho proposed a translation invariant procedure that essentially consists of averaging out several de-noised versions of the signal (using orthogonal wavelets), obtained for several shifts, after un-shifting (Coifman & Donoho, 1995). In simple terms, the procedure consists of performing the sequence of operations "Average[Shift – De-noise – Unshift]", a scheme named as "Cycle Spinning" by Coifman. With such a procedure, not only the pseudo-Gibbs phenomenon near the vicinities of discontinuities is greatly reduced, but also the results are often so good that lower sampling rates can be employed.

The choice of a proper thresholding criterion was also the target of various contributions, and several alternative approaches have been proposed, such as those based on cross-validation (Nason, 1996), minimum description length (Cohen *et al.*,

1999), minimization of Bayes risk (Ruggeri & Vidakovic, 1999) and on level-adaptive Baysean modelling in the wavelet domain (Vidakovic & Ruggeri, 2001). More elaborate discussions regarding this topic can be found elsewhere (Jansen, 2001; Nason, 1995a). The simultaneous choice of the decomposition level and/or wavelet filter was addressed by Pasti *et al.* (1999) and Tewfik, Sinha & Jorgensen (1992).

Image de-noising does not encompass any fundamental difference from 1D signal de-noising, apart from the fact that a 2D wavelet transform is now required. The computation of the 2D wavelet transform can be implemented by successively applying 1D orthogonal wavelets to the rows and columns of the matrix of pixel intensities, in an alternate fashion, implicitly giving rise to separable 2D wavelet basis (tensor products of the 1D basis functions). Non-separable 2D functions are also available (Jansen, 2001; Mallat, 1998; Vetterli & Kovačević, 1995).

Both in the context of 1D or 2D data analysis, an extension of the wavelet transform is often used, called Wavelet Packets. Wavelet packets provide a library or dictionary of basis sets for a given wavelet transform, built by successively decomposing not only the approximation signals at increasingly coarser scales using the high-pass and low-pass filtering operations followed by dyadic downsampling (as happens with the orthogonal wavelet transform), but also the details signals that are obtained along with them. As a result, there are now $2^{N/2}$ different basis sets for a signal of length N (Mallat, 1998), whose basis functions cover the whole time/frequency plane in a much more flexible way, from which derives the potential for generating more efficient representations for signals with complex behaviours in this plane, than those obtained with orthogonal wavelets: not only do we have more freedom in how to cover the time/frequency plane using different arrangements of "tiles",[11] but, furthermore, we are now able to select the best "tiling" for a specific application. Therefore, in order to choose an adequate basis set for a given application, efficient algorithms were developed to find the one that optimizes given quality criteria, such as entropy minimization (Coifman &

[11] By a "tile" we mean the area in the time/frequency space where a function of the basis set concentrates a significant fraction of its energy. It is also often called an "Heisenberg box" (Hubbard, 1998; Mallat, 1998).

Wickerhauser, 1992; Walczak & Massart, 1997b). This added flexibility does not come however without a cost, since the computational complexity is no longer of $O(N)$, as happens with the orthogonal wavelet transforms, but of $O(N\log_2(N))$ (Burrus *et al.*, 1998; Coifman & Wickerhauser, 1992; Mallat, 1998, 1999).

The approaches referred so far consist of implementing de-noising schemes through *off-line* data processing. Within the scope of *on-line* data rectification, where the goal is also the accommodation of errors present in measurements in order to improve data quality for accomplish other tasks, such as process control, process monitoring and fault diagnosis, Nounou and Bakshi (1999) proposed a multiscale approach for situations where no knowledge regarding the underlying process model is available. It basically consists of implementing the classical de-noising algorithm with a boundary corrected filter in a sliding window of dyadic length, retaining only the last point of the reconstructed signal for on-line use (On-Line Multiscale rectification, OLMS). When there is some degree of correlation between the different variables acquired, Bakshi, Bansal & Nounou (1997) presented a methodology where PCA (Appendix D) is used to build up an empirical model for handling such redundancies, and, finally, for the situation where our knowledge about the systems structure is deep enough such that a linear dynamical state-space model can be advocated for the finest scale behaviour, a multiscale data rectification approach was also proposed, using a Baysean error-in-variables formalism (Ungarala & Bakshi, 2000).

Other examples regarding applications of wavelets de-noising procedures in Chemical Engineering and related fields, include noise removal from industrial data (Nesic *et al.*, 1997), spectra (Chau *et al.*, 2004; Leung *et al.*, 1998a) such as near-infrared (NIR) (Depczynsky *et al.*, 1999; Walczak & Massart, 1997a), as well as from data generated in other analytical devices, namely, in electrochemistry, chromatography and capillary electrophoresis (Chau *et al*., 2004). Doymaz *et al.* (2001) have addressed the issue of filtering process signals also corrupted with outliers besides noise, proposing a wavelet based robust methodology.

2.2 Signal and Image Compression

The main goal in data compression is to represent the original signal (audio, image or even a sequence of images, i.e., a film) with as few bits as possible without compromising its end use. The usual steps involved in a ("lossy") compression scheme are as follows:

Signal Transformation → Quantization → Entropy Coding

First the signal is expanded into a suitable basis set (i.e., transformed); then, the expansion coefficients (i.e., the transform) are mapped into a countable discrete alphabet; finally, another map is used, where a new arrangement of coefficients is built up, such that the average number of bits/symbol is minimized. The first and the last steps are reversible, so that we can move forward and backward without losing any information, but the second stage (quantization) involves approximation, and, therefore, once we have gone through it, we can no longer recover the original coefficients (during decompression). It is precisely in this second step that many of the small wavelet coefficients are usually set to zero, and a high percentage of the compression arises from dropping out many wavelet coefficients in this way (Strang & Nguyen, 1997; Vetterli & Kovačević, 1995). This approach is being currently used, for instance, by the FBI services to store fingerprints with compression ratios of the order of 15:1, using wavelet transformation (linear phase 9/7 filter) together with scalar quantization (Strang & Nguyen, 1997). A wavelet based compression scheme is also adopted in the JPEG 2000 compression spec, with compression levels up to 200:1 being obtained for images in the "lossy" mode (where the potential of using wavelets can be fully used; a 9/7 wavelet filter is adopted) while the typical 2:1 compression ratio is achieved in the "lossless" mode (i.e., without the quantization step, using a 5/3 wavelet filter).

To get some practical insight into the compressing potential underlying wavelet-based compression, a fingerprint digitized image is presented in Figure 2.2 (a), as well as another version of it where only 5% of the original wavelet packet coefficients were retained (b), with all the remaining ones set equal to zero (a basis was selected using the best-basis algorithm of Coifman and Wickerhauser, as implemented in the WaveLab toolbox). As can be verified by comparing the two images, even though a high percentage of coefficients is being eliminated, the quality of the image remains quite satisfactory and close to the original.

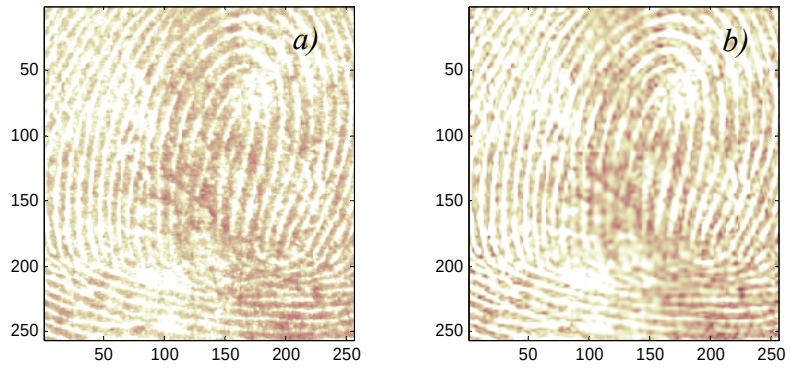


Figure 2.2. Original digitized fingerprint image (a) and a compressed version of it where 95% of the wavelet packet coefficients were set equal to zero (b).

Some examples of data compression applications in Chemical Engineering and related fields include the issue of on-line (Bakshi & Stephanopoulos, 1996; Misra *et al.*, 2001; Misra *et al.*, 2000; Trygg *et al.*, 2001) and off-line (Nesic *et al.*, 1997) industrial data compression. More examples can be found elsewhere (Chau *et al.*, 2004; Depczynsky *et al.*, 1999; Leung *et al.*, 1998b; Staszewski, 1998; Walczak & Massart, 1997a, 1997b; Walczak & Massart, 2001).

2.3 Regression Analysis

Wavelets have been applied both in parametric as well as in non-parametric regression analysis. Applications in parametric regression analysis usually involve compression of the predictor space when it presents serial redundancy, i.e., when there is a functional relationship linking the values of successive variables, as is the case when these are relative to wavelengths from digitized spectra, a common situation in multivariate calibration. By eliminating components with low predictive power, it is possible to reduce the variability of predictions (Alsberg *et al.*, 1998; Cocchi *et al.*, 2003; Depczynsky *et al.*, 1999; Eriksson *et al.*, 2000; Jouan-Rimbaud *et al.*, 1997) and construct more parsimonious models (Alsberg *et al.*, 1998; Trygg, 2001; Trygg & Wold, 1998), i.e., models encompassing a lower number of predictor variables, without compromising prediction ability. Furthermore, the use of the wavelet transform brings to the analysis the concept of *scale* and *characteristic frequency bands*, adding a new dimension to the regression models: *interpretation*. Therefore, not only the estimated

models may become *lighter* and equally or even more *effective*, but also easier to interpret, a feature explored by different authors (Alsberg *et al.*, 1998; Teppola & Minkkinen, 2000, 2001). Several strategies were proposed for selecting the number of transformed predictors (i.e., wavelet coefficients) to be included in the model, such as those based upon the variance spectra of the coefficients, where the ones with largest variance are selected (Trygg & Wold, 1998), leave-one-out cross-validation (Cocchi *et al.*, 2003), root mean square error (RMS), truncation of elements in the PLS weight vector followed by re-orthogonalization and mutual information (Alsberg *et al.*, 1998).

Alsberg *et al.* (1998) also refer the interesting relationship between the regression vector for the linear model involving the original variables, b, and that for the wavelet transformed variables, b_w, which, for the situation where only input spectra are wavelet transformed, is simply $b_w = Wb$ (meaning that b_w is the wavelet transform of b). This result also holds for PLS, but the equivalency is destroyed if the wavelet coefficients are processed (e.g. subject to some thresholding operation), as is often the case.

Multiscale PLS, a modelling framework consisting of estimating PLS models at each scale to capture the relationship between wavelet coefficients of predictors and response (the final prediction is obtained upon application of the wavelet reconstruction formula), is briefly addressed by Teppola & Minkkinen (2001), who also report some unsolved problems in this field.

Wavelet non-parametric regression methodologies present many resemblances to wavelet thresholding de-noising methodologies (Nason, 1994, 1995a, 1995b). Zhang (1995) presented an approach that combines elements from non-parametric and parametric regression for addressing the situation of moderately large input dimensionality.

A related topic regards the estimation of probability density functions, and several wavelet density estimators have been developed (Herrick, 2000; Safavi *et al.*, 1997), some of them also applied in Chemical Engineering applications, namely for process monitoring (Safavi *et al.*, 1997). Walter (1994) points out that a density estimator also leads to an estimator of the non-parametric regression function, $r(x) = E\left(Y \mid X = x\right)$.

2.4 Classification and Clustering

Classification and clustering constitute a final stage in the implementation of pattern recognition. Pattern recognition can be briefly described as a succession of transformations from the measurement space, *M*, to the feature space, *F*, and, finally to the decision space, *D*, through which an object is classified or clustered, after being properly measured and its relevant features for decision retained (Pal & Mitra, 2004):

$$M \rightarrow F \rightarrow D$$

In a classification problem, the information regarding class labels, $d \in D$, for the objects belonging to the training set is available, and is used to develop a classifier (decision function) that predicts class memberships for new objects, which is symbolically represented by the application, $\delta : F \rightarrow D$. On the other hand, in clustering no such *a priori* knowledge exists, and the goal is to unravel the underlying *similarities* between the objects, grouping those whose features are similar in some sense (Theodoridis & Koutroumbas, 2003). For these reasons, classification is also sometimes referred to as supervised (machine) learning and clustering as unsupervised (machine) learning.

The feature selection/extraction stage, $\phi : M \rightarrow F$, plays a key role in the success of a classification or clustering methodology (Pal & Mitra, 2004), as it is during this phase that the most relevant features for decision purposes are retained, usually along with a significant dimensionality reduction (Pal & Mitra, 2004; Walczak *et al.*, 1996), as quite often a large portion of the information contained in raw measurements does not bring added value for the final decision to be made through mapping δ. In this context, wavelet transforms can be quite useful, given their ability to concentrate the underlying deterministic features immersed in the signal into a few high magnitude coefficients (energy compaction property), whereas uninformative stochastic disturbances are spread over all coefficients, which set the conditions for developing effective coefficient selection methodologies, such as, for instance, those based on cross-validation, prediction power of signals reconstructed at a given scale (Alsberg, 2000) and on discrimination measures regarding wavelet packet coefficients (Cocchi *et al.*, 2004; Cocchi *et al.*, 2001).

Therefore, wavelets have been integrated in the feature extraction stage of classification pattern recognition problems in Chemical Engineering and related fields, namely in

problems with NIR data (Vannucci *et al.*, 2005; Walczak *et al.*, 1996), HS-SPME/GC signals (Cocchi *et al.*, 2004), X-ray diffractograms (Cocchi *et al.*, 2001), vibration analysis (Staszewski, 1998) and images (Theodoridis & Koutroumbas, 2003), as well as in clustering approaches involving spectra data from various sources (Alsberg, 1999; Donald *et al.*, 2005) and industrial data (Wang, 1999). In process operations, they have been applied to process operating region recognition (Zhao *et al.*, 2000) and to identify patterns in control charts (Al-Assaf, 2004).

2.5 Process Monitoring

The *energy compaction* and *decorrelation* properties associated with the wavelet-based multiscale representation of data provide an adequate way for effectively detecting undesirable events with a wide range of time/frequency location and localization patterns, as well as to incorporate the natural complexity of the underlying phenomena in process monitoring reference models. Therefore, a considerable number of approaches were already developed to explore such a potential (Ganesan *et al.*, 2004), as the following paragraphs attest. In this section we present a number of such works, beginning with MSSPC and related approaches in the following section, and then moving on to other monitoring methodologies based on alternative modelling formalisms, and finalizing with developments regarding the important case of profile monitoring.

2.5.1 Multiscale Statistical Process Control (MSSPC)

Addressing univariate SPC (USPC), Top and Bakshi (1998) proposed the idea of following the trends of wavelet coefficients at different scales using separate control charts. The state of the process is confirmed by reconstructing the signal back to the time domain, using only coefficients from scales where control limits were exceeded, and checking against a detection limit calculated from such scales (where significant events were detected). The approximate decorrelation ability of the wavelet transform makes this approach suitable even for autocorrelated processes, the signal power spectrum being accommodated by the scale-dependent nature of statistical limits. Furthermore, energy compaction enables the effective detection and extraction of underlying deterministic events. The multiscale nature of this framework lead the

authors to point out that it unifies Shewart, CUMSUM and EWMA procedures, as these control charts essentially differ in the scale at which they represent data (Bakshi, 1999).

Regarding multivariate applications, Kosanovich and Piovoso (1997) presented an approach where the Haar wavelet transform coefficients from filtered data (using a finite impulse response median hybrid filter) were used for estimating a PCA model, which is then applied for monitoring purposes, but it was with Bakshi (1998) that the first structured multivariate MSSPC methodology was established in the Chemical Engineering field. It is based on the so called multiscale principle components analysis (MSPCA), which combines the wavelet transform ability to approximately decorrelate autocorrelated processes and enable the detection of deterministic features present in the signal, together with the PCA ability to model the variables correlation structure (Appendix D). In MSPCA, PCA models are estimated for the wavelet coefficients at each scale, followed by a thresholding operation that separates the deterministic features from stochastic components of the signal, after which a PCA model in the original domain is estimated from the covariance matrix that combines the contributions from those scales where thresholding limits were violated (Figure 2.3).

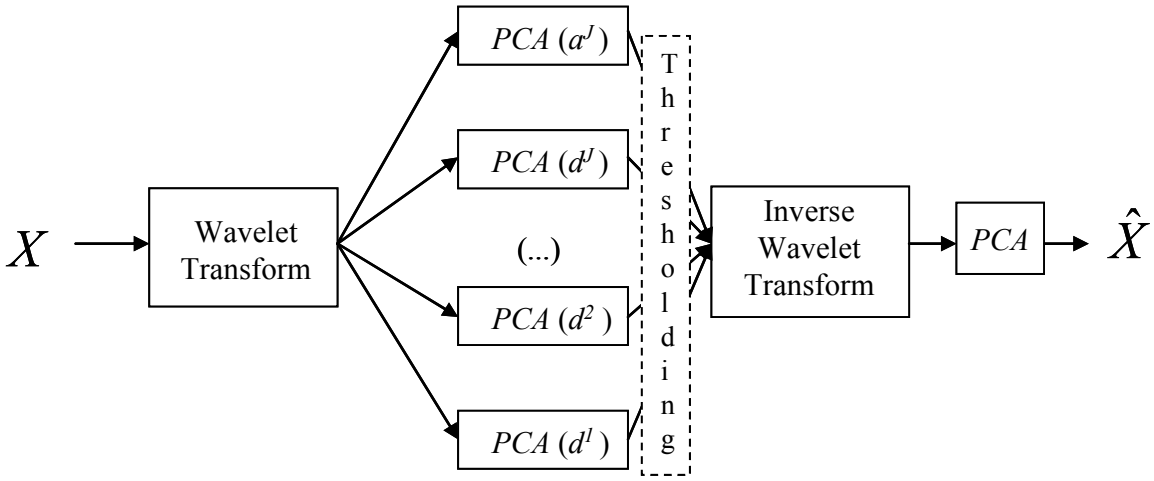


Figure 2.3. Schematic representation of the multiscale principal components analysis (MSPCA) methodology (Bakshi, 1998).

This methodology was applied to process monitoring, with the PCA models computed independently at each scale being used to implement separate PCA-MSPC control

charts (Bakshi, 1998). Once again, the scales where significant activity is detected are those that will be used to reconstruct the combined covariance matrix at the finest scale, through a scale-recursive procedure, in order to perform the final test that confirms or refutes the occurrence of abnormal perturbations detected at any scale(s). Following the denomination established by the author, this method will be here referred as multiscale statistical process control (MSSPC). A theoretical analysis of the properties underlying MSSPC can be found elsewhere (Aradhye *et al.*, 2003).

Several other works report improvements or modifications made to the original base procedure. In Kano *et al.* (2002), the monitoring procedure based upon MSPCA is integrated with methodologies designed for detecting changes in the correlation structure of data and in their distribution.

Misra *et al.* (2002) propose using MSPCA with variable grouping and the analysis of contribution plots whenever a significant event is detected in the control charts at any scale, in order to monitor the process, and, simultaneously, perform early fault diagnosis.

Yoon & Macgregor (2001, 2004), on the other hand, developed an approach based on a multiscale representation of data in the *original time domain* (i.e., not in the wavelet transform domain) that encompasses the successive extraction of principal components for an "extended set" (all variables represented at all scales), according to the decreasing magnitude of eigenvalues for the associated covariance matrix. It turns out that, owing to the orthogonal properties of the wavelet transform, the loadings obtained through this procedure only contain non-zero entries for variables represented at the same scale, and, therefore, each extracted component strictly conveys information regarding a specific scale. As such, results are not very different from what can be achieved with the classical MSSPC for the same number of principal components. However, this approach allows for ranking the relevant structures underlying overall data variability, in terms of the contributions from the variables covariance at different scales, and provides an hierarchic way of performing fault diagnosis, by first identifying the most relevant scales for a given fault detected by the MSSPC statistics (T^2 or the squared prediction error, SPE) and then looking to the variable contributions at that scale.

Rosen (2001) presented a methodology, also based on a multiscale representation in the original time domain, where the components from different scales are combined using

background knowledge available about the process, in order to reduce the number of monitoring statistics available when all the scales are monitored separately, and to provide physical insight to the scales under analysis. In this approach, there is no reconstruction stage, as happens in Bakshi's (1998) MSSPC, and the coarser scale coefficients are omitted from the monitoring procedure, to allow for adaptation (in the mean) to non-stationary data.

State space modelling based on Canonical Variate Analysis (CVA) was also adopted instead of PCA to handle variable cross-correlation, as well as dynamics (Alawi *et al.*, 2005). As the states and residuals still present autocorrelation, the authors used MSPCA to monitor them, leading to improved performance from the standpoint of the *detection delay/false alarm rate* balance, regarding alternative approaches based on CVA without using the wavelet based methodology, but based on theoretical control limits derived under the assumption of serially independent residuals, and the one where limits are calculated from the Empirical Reference Distribution (ERD). However, no comparison regarding the base MSSPC methodology is provided by the authors.

2.5.2 Alternative Multiscale Monitoring Approaches

Other multiscale approaches to process monitoring have also been developed, whose nature is quite distinct from MSSPC, and therefore referred in this separate subsection. They are related with alternative modelling paradigms, such as non-linear black-box modelling or hidden Markov trees.

Multiscale monitoring approaches using non-linear black-box modelling for building a reference process model were developed, such as those based on non-linear PCA, where an IT-net (input-training neural network) was adopted for establishing the non-linear mapping (Fourie & de Vaal, 2000; Li & Qian, 2004; Shao *et al.*, 1999).

The wavelet figure extraction ability has also been integrated with ART (adaptive resonance theory) frameworks in approaches for identifying operational states (Chen *et al.*, 1999; Li *et al.*, 2004; Wang, 1999; Wang *et al.*, 1999) and Maulud *et al.* (2005) proposed a bi-scale monitoring approach applied to the original domain, where an orthogonal non-linear PCA mapping is adopted for following the low frequency scales, and linear PCA to monitor the highest frequency bands. The authors also presented a graphical method for selecting the decomposition depth, based on the explained

variance captured by the PCA models derived upon neglecting successively detail coefficients below an increasing scale index.

Aradhye *et al.* proposed a multiscale fault detection strategy based on the ART-2 framework, where this clustering mechanism is used at each scale to select the significant wavelet coefficients for reconstruction (in the scale selection layer), and also after the reconstruction stage, where the cluster prototypes generated in the training phase are employed to classify new incoming observations (diagnosis layer) (Aradhye *et al.*, 2004; Aradhye *et al.*, 2002). The diagnosis layer is composed by $2^{J_{dec}+1}$ ART-2 networks (J_{dec} is the decomposition depth of the wavelet transform), i.e., one per each scale combination that can be obtained in the reconstruction stage, so that there is always an adequate prototype for each reconstructed signal (i.e., relative to the same selected scales). Despite the difference in the tools involved, there is a certain structural similarity between this approach and Bakshi's (1998) MSSPC, namely the existence of a figure extraction phase in the wavelet domain (scale selection layer), with the final decision being made on the reconstructed domain, based only upon those scales where significant events were detected (diagnosis layer).

Wavelets were also integrated in neural networks frameworks as activation functions, to perform fault diagnosis in dynamical systems (Zhao *et al.*, 1998).

Bakhtazad *et al.* (2000) developed a framework for the detection and classification of abnormal situations using the so called multidimensional wavelet domain hidden Markov trees (a multidimensional hidden Markov tree built over the wavelet coefficients calculated from the scores of a PCA model estimated from pre-processed raw data). More examples of related approaches can be found in Crouse *et al.* (1998) and Sun *et al.* (2003).

Luo *et al.* developed methods for detecting faults in isolated sensors, by analysing the intermediate frequency band obtained by applying the wavelet transformation with a decomposition depth of three on non-overlapping moving data windows, using parametric (Luo *et al.*, 1999) and non-parametric (Luo *et al.*, 1998) statistical tests.

Teppola & Minkkinen (2000) have also employed the multiresolution decomposition in order to remove seasonal and low-frequency trends from signals, which are often detrimental for the prompt detection of small and moderate-level transient phenomena.

Further applications of wavelets in process monitoring and fault detection can be found elsewhere (Alexander & Gor, 1998; Daiguji *et al.*, 1997; Jiang *et al.*, 2000; Jiao *et al.*, 2004; Tsuge *et al.*, 2000; Watson *et al.*, 1999).

2.5.3 Multiscale Monitoring of Profiles

With the development of instrumentation technology, one has now frequently to deal with situations where data is organized is such a way that the object of monitoring consists of a whole array of values (e.g. spectra, images, batch profiles) or a relationship between the response and the explanatory variables instead of their univariate or multivariate distributions (Kang & Albin, 2000; Kim *et al.*, 2003; Woodall *et al.*, 2004). We will refer to this type of applications as *profile* monitoring problems, and in this section we focus on the first type of scenario (monitoring arrays of values), where multiscale approaches based on wavelets have also been proposed.

Trygg *et al.* (2001) applied a 2D wavelet transformation to compress data from NIR (near-infrared) spectra collected over time and estimated a PCA model for this 2D compressed matrix, which was then used to check whether new incoming spectra deviate from those collected during normal operation.

Wavelet applications can also be found in the context of image analysis to monitor paper quality issues, such as paper formation, i.e., the degree of uniformity in the fibre network that constitutes paper (Bouydain *et al.*, 1999) and printing quality (Bernié *et al.*, 2004).

Other applications to process monitoring of profiles based on wavelet coefficients and metrics derived from them, can be found in: quadropole mass spectrometry data from rapid thermal chemical vapour deposition process (Lada *et al.*, 2002), tonnage signals from a stamping process (Jin & Shi, 1999; Jin & Shi, 2001), analysis of the central azimuth curve of antenna signals (Jeong *et al.*, 2004), data acquired from a semi-batch copolymerization process (Zhao *et al.*, 2000) and electrochemical noise data (fluctuation in potential) to characterize localized corrosion processes (Dai *et al.*, 2000).

2.6 System Identification, Optimal Estimation and Control

2.6.1 System Identification and Optimal Estimation

System identification is *"(...) the determination, on the basis of input and output, of a system within a specified class of systems, to which the system under test is equivalent.*" (L. A. Zadeh, in Åstrom & Eykhoff, 1971). It plays a central role in any application that requires adequate representations for input/output relationships (Ljung, 1999), namely optimal estimation, where the goal is now to figure what the true underlying value of a variable at a given time would be, using the information contained in a noisy realization (measurement corrupted with noise), over a finite time interval, say $[0, T]$. "Estimation" encompasses several problems – *prediction*, *filtering* and *smoothing* – according to the time instant where the value is to be estimated (Jazwinsky, 1970). Considering that we want to estimate $x(t)$ and are currently at time T, then we have the following kinds of problems, according to the location of t:

- $t > T$: prediction;
- $t = T$: filtering;
- $t < T$: smoothing.

There are many ways in which wavelets can be used for system identification and their application scenarios range from time-invariant systems (Kosanovich *et al.*, 1995; Pati *et al.*, 1993) to non-linear black-box modelling, for instance in the identification of Hammerstein model structures (Hasiewicz, 1999) or in neural networks, as activation functions (Section 2.6.2).

Noticing that all standard linear-in-parameter system identification methods can be understood as projections onto a given basis set, Carrier & Stephanopoulos (1998) applied wavelets basis sets in order to develop a system identification procedure with improved performance in estimating reduced-order models and non-linear systems, as well as systems corrupted with noise and disturbances, by focusing on the open-loop cross-over frequency region. Plavajjhala *et al.* (1996) used wavelet-based prefilters for system identification, proposing the parameter estimates computed at the scale (frequency band) where S/N ratio is maximal. The use of wavelets as basis functions is

also adopted by Tsatsanis & Giannakis (1993) in the identification of time-varying systems and a similar approach was followed by Doroslovački & Fan (1996) for adaptive filtering purposes, with the robustness issues being treated elsewhere (Doroslovački *et al.*, 1998).

Nikolaou *et al.* presented a methodology for estimating finite impulse response models (FIR) by compressing the Kernel (sequence of coefficients in the FIR model) using a wavelet expansion (Nikolaou & Vuthandam, 1998), and applied the same reasoning to nonlinear model structures, namely to quadratic discrete Volterra models (Nikolaou & Mantha, 1998).

Dijkerman & Mazumdar (1994) analysed the correlation structure of the wavelet coefficients computed for stochastic processes (see also Tewfik, 1992), and proposed multiresolution stochastic models as approximations to these original processes, motivated by the tree-based models (Bassevile *et al.*, 1992a), to be referred in Section 2.6.3.

Regarding multiscale optimal estimation, Chui & Chen (1999) implemented an on-line Kalman filtering approach that estimates the wavelet coefficients at each scale, and claimed evidence that it conducts to improved performance over the classical way of implementing it, when applied to a Brownian random walk process.

Renaud *et al.* (2005) proposed a procedure for multiscale autoregressive time series prediction (see also Renaud *et al.*, 2003), based on the redundant *à trous* wavelet transform (non-decimated Haar filter bank), and also developed a filtering scheme that takes advantage of such a decomposition, which is similar to Kalman filtering.

Other applications of wavelets have also been proposed in the field of time-series analyses, such as: estimation of parameters that define long range dependency (Abry *et al.*, 1998; Percival & Walden, 2000; Veitch & Abry, 1999; Whitcher, 2004); analysis of 1/*f* processes (Percival & Walden, 2000; Wornell, 1990; Wornell & Oppenheim, 1992), including fractional Brownian motion (Flandrin, 1989, 1992; Masry, 1993; Ramanathan & Zeitouni, 1991) as well the detection of 1/*f* noise in the presence of analytical signals (Mittermayr *et al.*, 1999); scale-wise decomposition of the sample variance of a time series (wavelet-based analysis of variance; Percival & Walden, 2000); analysis of electrochemical noise measurements in corrosion processes (Wharton *et al.*, 2003).

2.6.2 Wavelets and Neural Networks

Wavelets have also been used together with neural networks, in order to allow for a good description of the relationship between inputs and outputs in terms of both local and global approximation properties, by learning at multiple resolutions. Learning at multiple resolutions is a very useful feature, as data is often nonuniformly distributed in the input space, with some sparse regions where only a coarse description can be estimated, and with other more dense regions, where higher resolution mappings can be established (Bakshi & Stephanopoulos, 1993). It was in this context that the Wavelet Network, or Wave-Net was developed (Bakshi & Stephanopoulos, 1993; Zhang & Benveniste, 1992), were the role of the activation functions in the neural networks is played by wavelets. These networks retain all the advantages presented in those with localized learning (such as the Radial Basis Function Networks, RBFN), and add some more, namely regarding the ability to learn at multiple resolutions in an hierarchical way, from coarse to fine approximations, until the desired level of trade-off between accuracy and generalization has been reached, as well as enabling some interpretation of the mapping, estimating prediction errors and efficient training and adaptation (Bakshi & Stephanopoulos, 1993). More details on these approaches can be found elsewhere (Bakshi *et al.*, 1994; Juditsky *et al.*, 1994; Sjöberg *et al.*, 1995), as well as applications to modelling and optimization of an experimental distillation column (Safavi & Romagnoli, 1997), approximating single-input-single-output (SISO) and nonlinear second order processes (Oussar *et al.*, 1998).

Pati & Krishnaprasad (1993) have also proposed a wavelet network structure related to the ones referred in the previous paragraph, which was later applied in the field of analytical chemistry by Zhong *et al.* (2001), and Liu *et al.* (2000a) developed a wavelet-based network identification scheme for nonlinear dynamical systems.

Zhao *et al.* (1998) proposed an approach for dealing with the multidimensional case, where the wavelet networks approaches run into difficulties, given the high number of wavelet basis that must be considered, which grows exponentially with the number of dimensions, being $2^d - 1$ for the case where there are d inputs (Bakshi *et al.*, 1994) and also due to the associated numerical convergence issues. To overcome such limitations, the authors introduced a new multidimensional non-product wavelet function and proved its approximation ability.

2.6.3 Multiscale Modelling, Control and Optimal Estimation on Trees

Bassevile *et al.* (1992) established the foundations of a new theory for multiresolution stochastic modelling, along with developments regarding the associated techniques of optimal multiscale statistical signal/image processing. This topic will be described here in some detail, as it is on the basis of, and provided motivation to, other posterior developments. It is therefore important to get some insight into its methodology and underlying reasoning, in order to better understand subsequent applications in rather different fields. The methodology basically consists on studying stochastic processes indexed by nodes on *homogeneous trees*, in which different depths in the tree correspond to different scales in the signal or image representation.[12] In this approach, more than just an analytical tool, the wavelet transform does suggest the mechanism according to which the system evolves across the scales, occupying, in this sense, a similar position to the Fourier transform regarding stationary stochastic processes in the time domain, which made it so important in the analysis of such a class of systems, since it greatly simplifies their description, by providing, in particular, a way of transforming the process in a set of statistically uncorrelated frequency components (*whitening* the signal).[13]

The reasoning behind consists essentially in looking to the wavelet reconstruction procedure as defining a dynamic relationship that evolves across the scales, and where successive details, $d^j(\cdot)$, are added to a coarser representation at scale j in order to produce another with higher resolution at scale $j-1$ (the scale index $j-1$ indicates a signal representation with a finer resolution than the one indexed by j, as it integrates the additional detail information):

$$a^{j-1}(n) = \sum_k h(n-2k)\,a^j(k) + \sum_k g(n-2k)\,d^j(k) \tag{2.4}$$

[12] Homogeneous trees are an infinite, acyclic, undirected, connected graph that has exactly $q+1$ branches to other neighbouring nodes if its multiplicity is *q*.

[13] In fact, the Karhunen-Loève expansion for the covariance function of the multiscale process at a given scale consists of the wavelet transform (Chou, 1991), which clearly indicates the adequacy of the wavelet framework in the description of such systems.

For instance, if $d^j(\cdot)$ can be considered a white noise sequence, than equation (2.4) constitutes a first-order autoregressive model in scale, driven by white noise, but higher order models can be established, to describe a broader class of multiscale stochastic processes that found application in fields such as sensor data fusion. The tree-based topology, where these systems evolve, can be rather involved, but in the simplest case it reduces to a dyadic tree (an homogeneous tree with multiplicity $q = 2$), which is naturally associated with the Haar wavelet transform. Figure 2.4 presents such a tree, $\mathcal{T}$, that comprises all nodes, *t*, indexed by the ordered pair (*scale index*, *shift index*) along with the operators that define moves on $\mathcal{T}$, necessary to specify the local dynamics, namely:

α – Left forward shift;

β – Right forward shift;

δ – Interchange operator (move to the nearest point in the same horocycle[14]);

$\overline{\boldsymbol{\gamma}}$ - Backward shift.

The ***0*** operator should also be added to this set, representing the identity operator (no move). The convention is that the *left-most* operator is applied first, e.g. $t\alpha\overline{\gamma} = t\beta\overline{\gamma} = t$, (where *t* is a node).

[14] Nodes at the same *distance* from the *boundary point* are said to belong to the same *horocycle* (in other words, they consist of points at the same horizontal level). A boundary point (denoted by -∞) must be selected in a dyadic tree, and, in practical applications, corresponds to the root node (Figure 2.4). The distance between two nodes, t_1 and t_2, $d(t_1, t_2)$ is defined in this context as the number of nodes in the shortest path linking them (Bassevile *et al.*, 1992a; Stephanopoulos *et al.*, 1997b).

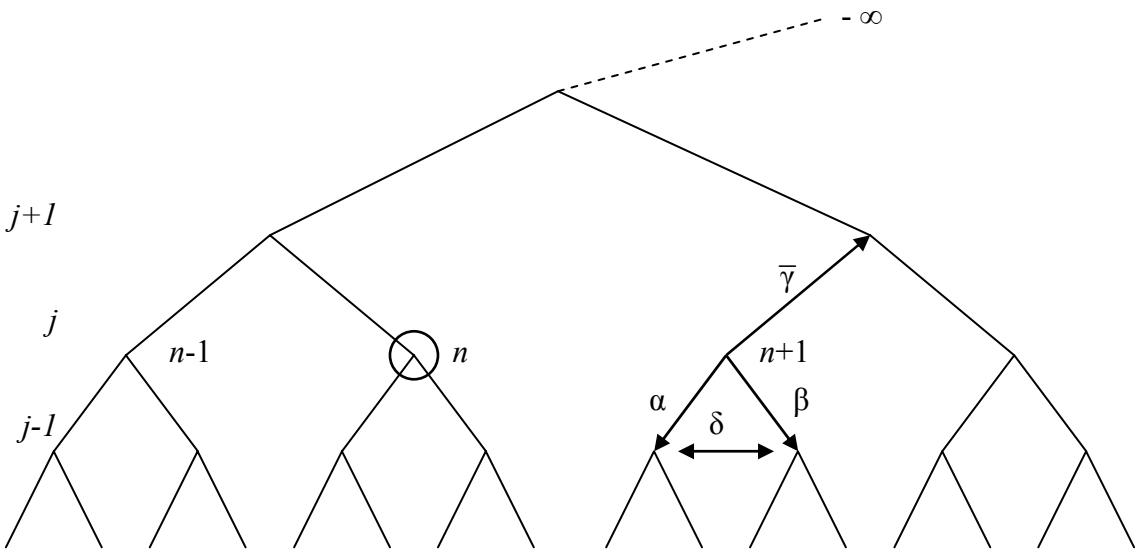


Figure 2.4. Dyadic tree, in which to each horizontal level (or horocycle) corresponds a scale index (*j*-1, *j*,…), with the nodes being completely defined by adding another index relative to their horizontal position (the shift index). Therefore, the pair (*scale index*, *shift index*), given by (j,n), completely defines the node signalled by a circle in the figure. Also presented are the translation operators that are used to move from one node to another one located in its neighbourhood, and are instrumental to write down the equations for the dynamical recursions in scale, that define multiscale systems.

After adapting the necessary fundamental theoretical notions to establish a systems theory on trees, such as the distance between two nodes, the authors developed *rational system functions* that would also be *causal in scale* and present an equivalent property, in scale, to the *shift-invariance* or *stationarity*, in time. The theory resulted in the parameterization of multiscale autoregressive models, such as $y_t = a y_{t\bar{\gamma}} + \sigma W_t$ (first-order autoregressive model in scale), as well as to the estimation of stationary processes and state models on dyadic trees. The topic of multiscale modelling and optimal estimation was further explored by the research group lead by Professor Alan S. Willsky at MIT (Bassevile *et al.*, 1992b, 1992c; Benveniste *et al.*, 1994; Chou, 1991; Chou *et al.*, 1993; Chou *et al.*, 1994a; Chou *et al.*, 1994b; Daoudi *et al.*, 1999; Golden, 1991; Ho, 1998; Luettgen *et al.*, 1993), and was revised in Willsky (2002), where an extended list of applications is presented.

Claus (1993) analysed the identification of multiscale autoregressive models (MAR) from a sample signal. Fieguth & Willsky (1996) used multiscale models on trees to

estimate the Hurst parameter of fractional Brownian motion, using a maximum likelihood approach.

Stephanopoulos *et al.* (1997b) also explored models defined on dyadic or higher-order homogeneous trees, whose nodes are used to index the values of any variable associated with the state and output equations (states, inputs, ouputs, modelling errors and measurement errors). These multiscale models on trees are *entirely consistent* with their time domain counterparts, since the equations linking the nodes were derived from linear, time-invariant models in the time domain, but their values now carry information localized in the hybrid domain of time/scale (or range of frequencies). For instance, the following homogeneous linear system

$$x(k+1) = \mathbf{A}x(k) \tag{2.5}$$

which, using the double indexing scheme can be written at the finest scale ($j = 0$) as

$$x(0,k+1) = \mathbf{A}x(0,k) \tag{2.6}$$

gives rise to the following multiscale description, based upon the Haar wavelet transform

$$x(t) = \mathbf{A}_{\alpha}^{(m)} x(t\overline{\alpha}) + \mathbf{A}_{\beta}^{(m)} x(t\overline{\beta}) \tag{2.7}$$

where

$$\begin{aligned} \mathbf{A}_{\alpha}^{(m)} &= \sqrt{2}\left(\mathbf{I} + \mathbf{A}^{2^m}\right)^{-1} \\ \mathbf{A}_{\beta}^{(m)} &= \sqrt{2}\left(\mathbf{I} + \mathbf{A}^{2^m}\right)^{-1} \mathbf{A}^{2^m} \end{aligned} \tag{2.8}$$

with $\overline{\alpha}$ and $\overline{\beta}$ being redefinitions of the backward shift ($\overline{\gamma}$), that specify the type of upward movement to do, if allowed ($\overline{\alpha}$ indicates a move to the parent node through a left-up shift, while $\overline{\beta}$ does the same through a right-up shift, as illustrated in Figure 2.5).

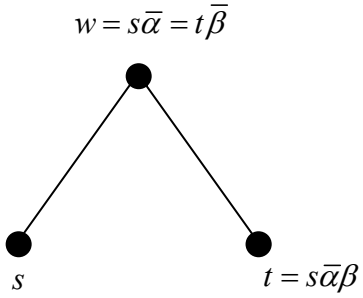


Figure 2.5. Illustration of operators for the upward moves: $\bar{\alpha}$ and $\bar{\beta}$.

The equations were written for the approximation coefficients, but analogous equations can be derived for the detail coefficients. Furthermore, forced linear systems (i.e., systems with inputs) can also be adequately described under this framework, giving rise to the following type of model structures:

$$\begin{aligned} x(t\alpha) &= \mathbf{A}_{\alpha}^{(m)} x(t) - B^{(m)} u(t\alpha) \\ x(t\beta) &= \mathbf{A}_{\beta}^{(m)} x(t) + B^{(m)} u(t\alpha) \end{aligned} \tag{2.9}$$

Stephanopoulos *et al.* (1997b) also pointed out that, as a consequence of the "closure requirement", according to which, "*The values of the states and inputs on a homogeneous tree, evolving by [equation (2.9)], must achieve "closure", i.e., be equal, with the values of the states and inputs on the discrete-time domain, evolving by [equation (2.10)]*"

$$x\left(m, {k}/{2^m} + 1\right) = A^{2^m} x\left(m, {k}/{2^m}\right) + \left(I + A^{2^{m-1}}\right)\left(I + A^{2^{m-2}}\right)\cdots(I + A)\, Bu\left(m, {k}/{2^m}\right) \tag{2.10}$$

the following two-scale model structure,

$$\begin{aligned} x(t\alpha) &= \mathbf{A}_{\alpha}^{(m)} x(t) + B_{\alpha}^{(m)} u(t\alpha) \\ x(t\beta) &= \mathbf{A}_{\beta}^{(m)} x(t) + B_{\beta}^{(m)} u(t\beta) \end{aligned} \tag{2.11}$$

cannot give rise to a discrete-time, causal model at the finest scale of the form:

$$x(k+1) = \mathbf{A}x(k) + \mathbf{B}u(k) \tag{2.12}$$

at any resolution. Therefore, such a model structure can not be used to adequately describe the behaviour of causal systems, a class that encompasses many (if not most) of the relevant phenomena in Chemical Engineering.

The authors have also analysed *stability* conditions defined for the multiscale representation, proving that (2.9) is ℓ_p-stable in the Lyapunov sense when its discrete-time domain counterpart representation, that was in its origin, is also ℓ_p-stable in the Lyapunov sense, and the same type of conclusions hold for system's *controllability* and *observability*. These types of models were then applied to several process systems engineering tasks, such as simulation of linear dynamical systems, multiscale optimal control and model predictive control (MPC), as well as state estimation with optimal fusion of measurements (further addressed in Dyer, 2000). The parallelizable nature of the computations performed by the algorithms developed was also highlighted by the authors. The issue of estimating the multiscale model structure is referred in Stephanopoulos *et al.* (1997a).

The topic of multiscale MPC is further detailed in Stephanopoulos *et al.* (2000) and Karsligil (2000). Krishnan & Hoo (1999) also presented a multiscale MPC strategy based on dyadic homogeneous trees and applied it to a continuous process and to a batch reactor.

Ungarala & Bakshi (2000) also proposed a multiscale approach for linear dynamic data rectification that explores a tree-based model structure similar to that proposed by Stephanopoulos *et al.* (1997b).

2.7 Numerical Analysis

Within the scope of numerical analysis, wavelets have been used for the solution of systems of equations and differential equations (Louis *et al.*, 1997; Nikolaou & You, 1994; Santos *et al.*, 2003), namely regarding applications to chromatography (Liu *et al.*, 2000b), combustion (Prosser & Cant, 1998) and cooled reverse flow reactors (Bindal *et al.*, 2003). Mahadevan & Hoo (2000) proposed a wavelet-based model reduction strategy for distributed parameter systems that give rise to a finite low-order model still representing the systems multiscale behaviour.

Beylkin *et al.* (1991) address the issue of fast application of dense matrices (or integral operators) to vectors, using a class of numerical algorithms based upon wavelets.

Binder (2002) proposed a wavelet-based multiscale methodology for on-line scalable dynamic optimization, which strives to use all the available time allocated for

computation in order to come up with the best possible solution at the moment where it is required. It belongs to the class of algorithms known as “any time algorithms”, that typically progress by refining an initial solution approximation, improving its quality along the iterations, so that they can provide the user with a solution at *any time*, the solution approximation being better and better, the longer the procedure is allowed to proceed. In this case, the initial solution is based on a coarse problem approximation, and improved by successively adding details.

Regarding the implementation of PCA and PCR over large data sets composed of spectra data or hyperspectral images, Vogt & Tacke (2001) proposed a preliminary step of wavelet based compression in order to reduce computation time during SVD decomposition. A similar strategy was proposed before by Walczak & Massart (1997b), but focused on the optimization of data compression, rather than on the minimization of overall computation time.

Part III

Background Material

The ideal engineer is a composite ... He is not a scientist, he is not a mathematician, he is not a sociologist or a writer; but he may use the knowledge and techniques of any or all of these disciplines in solving engineering problems.

N.W. Dougherty (Civil Engineer), 1955

Chapter 3. Mathematical and Statistical Background

Some classes of tools are adopted more frequently throughout this thesis, to explore, analyse and unravel the potential useful structure present on industrial data. They may be either applied separately, under simplified contexts, or combined into integrated frameworks, in order to take advantage of their complementary strengths under more complex scenarios. In this chapter, we analyse those presenting special relevance within the scope of our work. In particular, we address the basic concepts underlying statistical process control, to which some developments in this thesis will be directed to (within the scope of process monitoring), and the specification of measurement uncertainties information, that plays an important role in the techniques to be presented in Part IV-A (namely regarding regression and monitoring in noisy environments).

Furthermore, latent variable models and wavelet theory are also presented, as they often provide the basis for developing adequate methodologies to handle the multivariate and multiscale nature of data. In this context, we will present the motivation and general structure of latent variable models, as well as some current approaches for estimating their parameters, and introduce wavelet theory basics and nomenclature, to facilitate the understanding of its application in the multiscale approaches to be addressed, in particular, in Part IV-B.

Additional details concerning particular tools inside these classes will be presented in the forthcoming chapters, whenever appropriate, and more elaborate discussions regarding the techniques are referred to the relevant published literature.

3.1 Statistical Process Control (SPC)

Statistical Process Control (SPC) strives for achieving process stability and improving capability through the reduction of variability (Montgomery, 2001). It comprises a collection of several problem-solving tools, the major representatives of which are known as the "magnificent seven", as follows:

- Histogram or steam-and-leaf display
- Pareto chart
- Cause-and-effect diagram
- Defect-concentration diagram
- Scatter diagram
- Check sheet
- Control chart

The control chart, in particular, is probably the most sophisticated and eventually the most powerful of them (Montgomery & Runger, 1999). It enables the verification, at each data acquisition time, of whether the process is operating where it is expectable to be under a given reference scenario, where only *chance* or *common causes* of variation are occurring in the process (i.e., it is *in statistical control*), or if some *assignable* or *special* cause has taken place, which needs to be promptly detected, so that its root cause can be found and corrective actions undertaken, in order to prevent more losses due to poor quality. Control charts essentially consist of plots where the values for one more quality characteristics (or statistics calculated from them) are plotted, and where reference lines are also drawn, delimiting normal operation condition (NOC) regions (Kresta *et al.*, 1991), where only common causes are active (sometimes additional reference lines may appear, such as the centre line, that represents the average value of the statistic under normal operation conditions). The parameters that define the NOC region are set by analysing historical data collected under normal operating conditions, and specifying a significance level or a multiplicative constant to be applied to the normal operation variability parameter. An example of a control chart is the celebrated Shewhart control chart, proposed by Walter S. Shewhart, while working at the Bell

Telephone Laboratories, in the early 1920s (Kenett & Zacks, 1998; Shewhart, 1931), which basically consists of a time series plot of the quality characteristic, along with a centre reference line (given by the mean of data collected under normal operation conditions, $Center\ Line = \hat{\mu}_X$) and two control limits, an upper control limit ($UCL = \hat{\mu}_X + k\hat{\sigma}_X$), and a lower control limit ($LCL = \hat{\mu}_X - k\hat{\sigma}_X$), that define the NOC region (k can be set directly, for instance equal to three, as in Figure 3.1, or by specifying the significance level to be associated with the control chart).

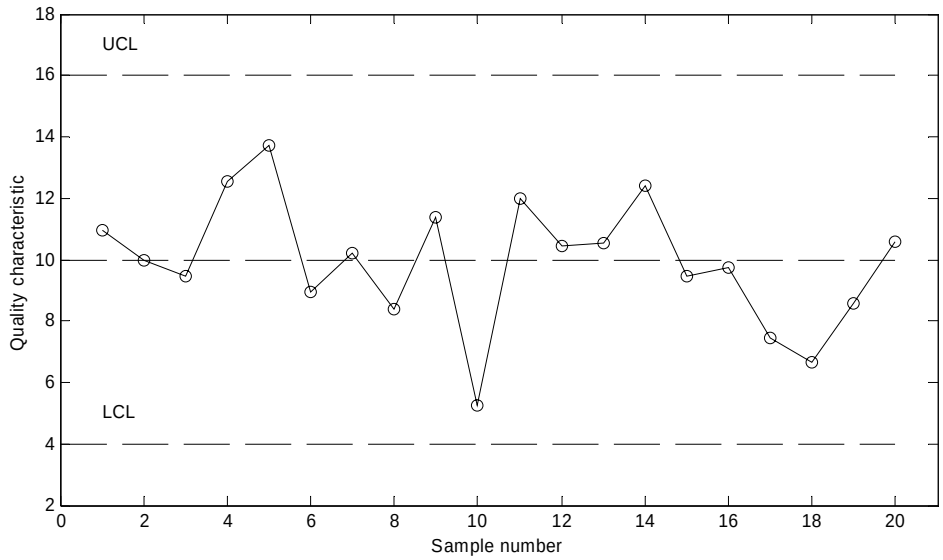


Figure 3.1. Example of a Shewhart control chart, with "three-sigma" control limits.

Other examples of control charts, with special sensitivity for detecting special events involving small shifts or slowly drifting processes, are the EWMA (Hunter, 1986) and the CUSUM control charts (Montgomery, 2001).

Until recently, there was a certain tradition of using control charts in the supervision of single isolated variables, usually referred to as *univariate* SPC (USPC) charts. However, it is well known that this procedure presents difficulties when dealing with multivariate data exhibiting correlated behaviour (Jackson, 1959; Kourti & MacGregor, 1995; MacGregor & Kourti, 1995; Montgomery, 2001; Tracy *et al.*, 1992). In fact, implementing several univariate SPC charts in parallel not only brings problems in the definition of the true overall significance level for the combined performance of the tests, but also erroneously makes the tacit assumption that a multivariate NOC region is an "hiper-rectangle", when, in fact, its shape is usually more similar to a

multidimensional ellipsoid. Therefore, there are sectors in the *assumed* NOC region that do not belong to the *actual* NOC region, which means that such a procedure will not detect certain special events. Thus, for those situations where a certain degree of correlation is present among the variables, Multivariate Statistical Process Control (MSPC) procedures were developed, and found to be more adequate for implementing SPC control chart procedures.[15] One example of such a procedure is the Hotelling T^2 control chart, for *i.i.d.* processes (independent and identical distributed) following a multivariate normal distribution (MacGregor & Kourti, 1995; Montgomery, 2001). Multivariate extensions for the CUSUM and EWMA control charts were also developed (MacGregor & Kourti, 1995).

As the number of variables increases, even the MSPC control chart methodology begins facing some difficulties: the effect of any change over any individual variable is diluted in the overall consideration of all variables contributing to the calculation of the statistic, considerably augmenting the time period required to detect a meaningful change in any one of them (Montgomery, 2001). Furthermore, the covariance matrix becomes nearly singular (MacGregor & Kourti, 1995), as more and more redundant information is conveyed by the set of correlated variables. The proposed approach to circumvent this problem is based on a latent variable description of the data structure, where the few underlying components, driving the variability exhibited by all the variables, are first extracted, and then focusing monitoring efforts over this reduced set of latent variables (or statistics calculated from them), as well as in the distance between each observation and its projection onto the lower dimensional subspace where such a reduced set lies (Jackson & Mudholkar, 1979; Kourti & MacGregor, 1995; Kresta *et al.*, 1991; MacGregor & Kourti, 1995; Wise & Gallagher, 1996). A more detailed description of this procedure will be provided in Section 3.3.

[15] In the remainder text, control chart SPC procedures will be simply referred to as SPC procedures, provided no confusion arises to other SPC tools.

3.2 Measurement Uncertainty

The quality of data is a key factor in data-driven analysis frameworks. On one hand, it depends on the data generating mechanism, that should be appropriate, according to the end use of collected data. For instance, it should be *maximally informative* when designing experiments for system identification purposes (Ljung, 1999), or faithfully represent normal operation when collecting data for process monitoring applications. On the other hand, it also depends on the signal to noise ratio (SNR) of the collected signals. In fact, any measured value has associated with it a certain uncertainty level, which should be specified in order to enable a sound use of data in the subsequent analysis. Establishing an analogy, the measurement value/measurement uncertainty pair acts like the two faces of a coin: both of them are necessary in order to have a valuable coin, otherwise it has very little value. If one face is lacking (measurement uncertainty), we may have a lot of numbers, being *numbers rich*, but of a limited value, as we do not really know what those numbers are really worth, and are therefore *information poor*.

Measurement noise features can adequately be specified within the scope of *measurement uncertainty*, which is a key concept in metrology, defined as a "*parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand* "[16] (ISO, 1993). Recommendations regarding the correct terminology and procedures to adopt in order to compute and specify measurement uncertainties can also be found in the text (ISO, 1993) "Guide to the Expression of Uncertainty in Measurement" (GUM), which was written following an initiative of the *Comité International des Poids et Mesures* (CIPM), that requested its executive body, the *Bureau International des Poids et Mesures* (BIPM), to address the problem of the expression of uncertainty in measurement, in conjunction with the national standard laboratories (see also Kessel, 2002; Kimothi, 2002; Lira, 2002).

According to GUM, the *standard uncertainty*, $u(x_i)$ (to which we will often refer simply as uncertainty), is expressed in terms of a standard deviation of the values

[16] Measurand is the "*particular quantity subject to measurement*" (ISO, 1993).

collected from a series of observations (the so called Type A evaluation), or through other adequate means (Type B evaluation), namely relying on an assumed probability density function based on the degree of belief that an event will occur. Numerical quantities, y, calculated from uncertain measurements, $\{x_i\}_{i=1:N}$, according to a functional relationship of the type

$$y = f\left(x_1, x_2, \cdots x_N\right) \tag{3.1}$$

turn out to be also uncertain quantities, and therefore should have associated an uncertainty value, the *combined standard uncertainty*, $u_c(y)$, which is calculated according to a propagation formula, such as the following:

$$u_c^2(y) = \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}u(x_i, x_j) = \sum_{i=1}^{N}\left(\frac{\partial f}{\partial x_i}\right)^2 u^2(x_i) + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{\partial f}{\partial x_i}\frac{\partial f}{\partial x_j}u(x_i, x_j) \tag{3.2}$$

Equation (3.2) is based on a Taylor series expansion neglecting second and higher order terms (Herrador *et al.*, 2005), which should be added when non-linearity becomes important. When the uncertainty is required to express an "*interval about the measurement result of a measurement that may be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the measurand*" (ISO; 1993), an adequate factor (the *coverage factor*) is chosen to multiply the standard uncertainty, in order to obtain the *expanded uncertainty*, $U = k_c u_c(y)$.

Acknowledging the importance of taking into account uncertainty information in data analysis, some authors have already directed their efforts towards the development of uncertainty-based approaches. Wentzell *et al.* (1997a) developed the so called maximum likelihood principal components analysis (MLPCA), which estimates a PCA model (Appendix D) in an optimal maximum likelihood sense, when data are affected by measurement errors exhibiting complex structures, such as cross-correlations along sample or variable dimensions. The reasoning underlying MLPCA was then applied to multivariate calibration (Wentzell *et al.*, 1997b), extending the consideration of measurement uncertainties to some input/output modelling approaches closely related to PCA. Bro *et al.* (2002) presented a general framework for integrating data uncertainties in the scope of (maximum likelihood) model estimation, comprehending MLPCA as a special case. The issue of (least squares) model estimation is also referred by Lira (2002), along with the presentation of general expressions for uncertainty propagation

in several input/output model structures. Both multivariate least-squares (MLS), and its univariate version, bivariate least-squares, (BLS), were applied in several contexts of linear regression modelling, when all variables are subject to measurement errors with known uncertainties (Martínez *et al.*, 2000; Río *et al.*, 2001; Riu & Rius, 1996). The issue of detecting analytical bias using a functional EIV model that incorporates uncertainty information was also addressed (De Castro *et al.*, 2004; Galea-Rojas *et al.*, 2003). On the other hand, Faber & Kowalski (1997) explicitly considered the influence of measurement errors in the calculation of confidence intervals for the parameters and predictions in PCR and PLS, and similar efforts can be found elsewhere (Faber, 2000; Faber & Bro, 2002; Phatak *et al.*, 1993; Pierna *et al.*, 2003).

3.3 Latent Variable Modelling

The increasing number of variables acquired in chemical processing units, associated with higher sampling rates, soon led to databases of considerable sizes, which gather huge amounts of records originated from several sources in the plant. Even the sole idea of what is a "large" data set have evolved during the last 40 years, at a rate of an order of magnitude per decade: if in the 70's a "large" data set was considered as such if it had more than 20 variables, nowadays a data set is considered to be "large" if it exceeds, 100 000 – 1 000 000 variables (Wold *et al.*, 2002). Therefore, techniques tailored to handle problems raised by the high dimensionality of data sets, along with the extensive use of computation power, are playing an increasingly important role in data analysis, and terms like "multivariate analysis" are being up-dated to "megavariate analysis" (Eriksson *et al.*, 2001). It is in this context that latent variable models gained considerable relevance, since they provide an adequate setting for developing approaches that can be very effective in different tasks, being furthermore quite often computationally efficient.

Large data sets usually contain redundancies (covariance) among different groups of variables. In other words, this means that the dimensionality of data sets (number of variables) is large when compared to the "true" dimensionality of the underlying process generating overall variability, i.e., the number of independent sources of variability that structures the overall dispersion of observed values. These independent sources of variability in industrial data are usually related to raw material variability,

process-related disturbances and other perturbations that might be introduced through other means, like operators interventions, and are usually of the order of magnitude of a dozen. Covariance between variables may have different origins:

- Dependencies caused by the underlying phenomena, such as conservation laws (mass and energy);
- Presence of control loops;
- Use of redundant instrumentation;
- The nature of the measuring devices employed. For instance, if the measurement devise is a spectrometer, the variables will be something like the "absorbance" or "reflectance" at a set of frequencies or wavelengths. In such a situation, the variables have a natural ordering (the frequency or wavelength), and the correlation arises from the spectral characteristics of the samples.

In these circumstances, a useful picture is provided by the latent variable model, that considers the process as being driven by a few, not observable (p) *latent variables*, t, with the m ($m \geq p$) measured variables, x, being the visible "outer" part of such an "inner" set of variation sources. Their mutual relationship is given by:

$$x = t\mathbf{P} + \varepsilon \tag{3.3}$$

where x and t are $1 \times m$ and $1 \times p$ row vectors, respectively, $\mathbf{P}$ is a $p \times m$ matrix of coefficients and ε is the $1 \times m$ row vector of random errors, that encompasses unstructured sources of variability such as measurement error, sampling error and unknown process disturbances (Burnham *et al.*, 1999). For n observations, the $n \times m$ data table, $\mathbf{X}$, that consists of n rows for the m different variables, can be written as,

$$\mathbf{X} = \mathbf{TP} + \mathbf{E} \tag{3.4}$$

where $\mathbf{T}$ is the $n \times p$ matrix of latent variables (each row corresponds to a different vector t, representing an observation of the p latent variables) and $\mathbf{E}$ is the $n \times m$ matrix that also results from stacking up the rows ε for each (multivariate) observation. Sometimes it is useful to separate the variables into two groups, $\mathbf{X}$ and $\mathbf{Y}$, for instance in order to use the model in future situations where the $\mathbf{Y}$ variables are not available or its prediction is required, under the knowledge of only the values for the variables from the $\mathbf{X}$-block. In such circumstances we can write (3.4) in the following form

$$\begin{aligned} \mathbf{X} &= \mathbf{TP} + \mathbf{E} \\ \mathbf{Y} &= \mathbf{TQ} + \mathbf{G} \end{aligned} \tag{3.5}$$

which can be directly obtained from (3.4), after a rearrangement with the variable grouping: $[\mathbf{XY}] = \mathbf{T}[\mathbf{PQ}] + [\mathbf{EF}]$. From (3.5) it is clear that there is no causality assumed between the variables belonging to the two blocks in the latent variable model. In fact, non-causality is also a characteristic of historical databases and normal operation data, situations where this model is often applied (MacGregor & Kourti, 1998). Blocks **X** and **Y** share a symmetrical role regarding the underlying latent variables, and their separation is only decided on the basis of the intended final use for the model.

Model (3.4) can be estimated using Factor Analysis and Principal Components Analysis (PCA, Appendix D), whereas (3.5) is often estimated using Principal Components Regression, PCR (Jackson, 1991; Martens & Naes, 1989) or Partial Least Squares, also known as Projection to Latent Structures, PLS (Geladi & Kowalski, 1986; Haaland & Thomas, 1988; Helland, 1988, 2001b; Höskuldsson, 1996; Jackson, 1991; Martens & Naes, 1989; Wold *et al.*, 2001). When the error structures are more complex, other techniques, like Maximum Likelihood Principal Components Analysis (MLPCA) for model (3.4), can be used (Wentzell *et al.*, 1997a).

Some useful features of these estimation techniques, besides handling the presence of collinearity in a natural and coherent way, are their ability to handle the presence of moderate amounts of missing data, by taking advantage of the existent correlation among variables (Kresta *et al.*, 1994; Nelson *et al.*, 1996; Walczak & Massart, 2001), their flexibility to cope with situations where there are more variables than observations, and the availability of diagnostic tools that portray information regarding the suitability of the models to explain the behaviour of new incoming data. For instance, in PLS it is possible to calculate the distance between the new observation in the **X**-space and its projection onto the latent variable space (space spanned by the latent variables, as defined by the **P** matrix), as well as to see whether this projection falls inside the domain where the model was estimated. These features enable checking whether a new observation is adequately described by the estimated latent variable model, and, furthermore, if it falls within the region used to build the model, therefore avoiding extrapolation problems in the prediction of variables belonging to the **Y** block. Another useful characteristic of these approaches is that several variables in the **Y** block can be

handled simultaneously. The flexibility and potential utility of the approaches based upon latent variables is illustrated in the following paragraphs, by mentioning several different application scenarios.

3.3.1 Process Monitoring

Process monitoring is a field where latent variable models have found generalized acceptance. Both PCA and PLS techniques have been extensively used as estimators of the structure underlying normal operation data. SPC based on PCA consists of using two statistics:

- T^2 *statistic* – to monitor variability within the PCA subspace, checking whether projections of new observations onto this subspace fall inside or outside the normal operation conditions region (NOC)

$$T_i^2 = t_i \Lambda^{-1} t_i^T \tag{3.6}$$

  where t_i is the i^{th} row of the score matrix, T, and Λ^{-1} is the inverse of the diagonal matrix with the p largest eigenvalues in descendent order of magnitude along the main diagonal;

- *Q statistic (also referred as SPE)* – to assess the adequacy of the principal components model in describing each new observation, by calculating the square distance for each new incoming observation to the principal components subspace

$$Q_i = e_i e_i^T \tag{3.7}$$

  where e_i is the i^{th} row of the residual matrix, $\mathbf{E}$.

Basically, Q is a "lack of model fit" statistic, while the Hotelling's T^2 is a measure of the variation within the PCA model. These two general monitoring features are normally present in any monitoring procedure based on a latent variable framework (Figure 3.2).

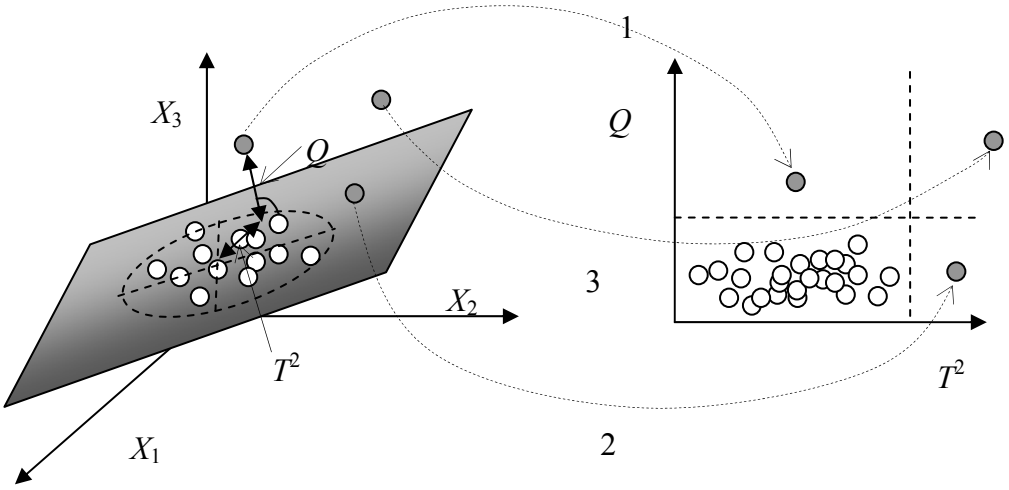


Figure 3.2. Illustration of a multivariate PCA monitoring scheme based on the Hotelling's T^2 and Q statistics: observation 1 falls outside the control limits of the Q statistic (the PCA model is not valid for this observation), despite its projection on the PC subspace falling inside the NOC region; observation 2, on the other hand, corresponds to an abnormal event in terms of its Mahalanobis distance to the centre of the reference data, but it still complies with the correlation structure of the variables, i.e., with the estimated model; observation 3 illustrates an abnormal event from the standpoint of both criteria.

Control limits for these statistics have been derived for the case of data following a multivariate normal distribution. The (upper) critical value for the T^2 statistic is:

$$T^2_{\lim}\left(p,n,\alpha\right)=\frac{p\left(n-1\right)}{\left(n-p\right)}F\left(p,n-p,\alpha\right) \tag{3.8}$$

where $F\left(\nu_1,\nu_2,\alpha\right)$ is the upper $\alpha\times100\%$ percentage point for the F distribution with ν_1 and ν_2 degrees of freedom (α is the chosen significance level). As for the Q statistic, the (upper) control limit is given by:

$$Q_{\lim}\left(p,\alpha\right)=\Theta_1\left(\frac{c_\alpha\sqrt{2\Theta_2 h_0{}^2}}{\Theta_1}+1+\frac{\Theta_2 h_0\left(h_0-1\right)}{\Theta_1{}^2}\right)^{\frac{1}{h_0}} \tag{3.9}$$

where

$$\Theta_i=\sum_{j=p+1}^{m}\lambda_j^i,\ i=1:3 \tag{3.10}$$

$$h_0 = 1 - \frac{2\Theta_1\Theta_3}{3\Theta_2^{\,2}} \tag{3.11}$$

and c_α is the upper $\alpha\times100\%$ percentage point for the standard normal distribution.

Individual scores can also be monitored using univariate SPC charts. Assuming that each score is *i.i.d.* with a zero mean normal distribution (which requires the data matrix to be previously centred at the mean vector calculated from reference data), the control limits for the i^{th} score are then given by:

$$t_{\lim}(i,\alpha) = \pm T(n-1,\alpha/2)\sqrt{\lambda_i} \tag{3.12}$$

where $T(n-1,\alpha/2)$ is the upper $\alpha/2\times100\%$ percentage point for the student's-t distribution with $(n-1)$ degrees of freedom.

In practice, some departure from the multivariate normal distribution is tolerated as the scores are themselves linear combinations of several variables, and thus shall conform more to the assumed Gaussian behaviour than individual variables, due to the Central Limit Theorem (CLT). As for the Q statistic, it is found in practice to be even more robust to departures from normality than the previous ones, because it is a "residual" statistic, measuring the unstructured variability present in data, after removal of the deterministic part through a PCA model.

Using only the T^2 and Q statistics, processes with dozens or hundreds of variables can be easily and effectively monitored (MacGregor & Kourti, 1995; Wise & Gallagher, 1996). Underlying this monitoring scheme is latent variable model (3.4), but we can also monitor processes using the latent variable model structure (3.5), estimated by PLS. In this case, the statistics adopted are usually the Hotelling's T^2 statistic applied to the latent variables and $SPE_y = \sum_{i=1}^{m_y}\left(y_{new,i} - \hat{y}_{new,i}\right)^2$, where $\hat{y}_{new,i}$ are the predicted values for the m_y **Y**-block variables in the i^{th} observation. When these variables are measured infrequently relatively to the acquisition rate for the **X**-block variables (as often happens with quality variables, with regard to process variables), then the statistic $SPE_x = \sum_{i=1}^{n}\left(x_{new,i} - \hat{x}_{new,i}\right)^2$ is used instead (or a modified version of it, that weights the variables according to their modelling power for **Y,** $SPE_x = \sum_{i=1}^{n} w_i\left(x_{new,i} - \hat{x}_{new,i}\right)^2$), where $\hat{x}_{new,i}$ is the projection of observation *i* in the latent variable subspace (Kresta *et*

al., 1991). The upper control limit for the Hotelling's T^2 statistic is calculated as in the case of SPC based on PCA, while the *SPE* statistic is based on a χ^2 approximation, with parameters determined by matching the moments of this distribution with those of the reference "in-control" data (Kourti & MacGregor, 1995; MacGregor *et al.*, 1994; Nomikos & MacGregor, 1995).

One important feature of latent variable frameworks in the context of process monitoring is the availability of informative diagnostic tools. The above referred statistics do detect abnormal situations effectively but do not provide any clue about what may have caused such behaviour. However, with the assistance of these diagnostic tools, one can give a step further towards reducing the number of potential root causes or even track down the source of the abnormal behaviour. This can be done through the use of contribution plots (Eriksson *et al.*, 2001; Kourti & MacGregor, 1996; MacGregor *et al.*, 1994; Westerhuis *et al.*, 2000), which basically are tools that "ask" the underlying latent variable model (estimated through PCA or PLS) about which are the variables that mostly contribute to unusual values of the monitoring statistics. For instance, in the case of SPC based on PCA, there are contribution plots available for each individual score, for the overall contribution of each variable to the Hotelling's T^2 statistic and for the contribution of each variable to the *Q* statistic. A hierarchical diagnostic procedure was also proposed by Kourti & MacGregor (1996), consisting of following first the behaviour of the Hotelling's T^2 statistic and, if an out-of-limit value is detected, checking the score plots for high values and then the variable contributions for each significant score.

The procedures mentioned so far are specially suited for continuous processes, where the assumption of a stationary mean and covariance structure holds as a good approximation of reality. If dynamics or autocorrelation are present in the variables, we can still adopt these procedures by expanding the original data matrix with time-lagged variables, in order to model both the cross-covariance and autocorrelation structures (Ku *et al.*, 1995; Ricker, 1988). Lakshminarayanan *et al.* (1997) describe an approach where dynamic modelling is introduced in the inner relationship of PLS using ARX or Hammerstein model structures, while Kaspar and Ray (1993) present an alternative procedure where a dynamic transformation is applied to the **X**-block time series, before applying PLS.

A special case of intrinsically dynamic processes are the so called batch processes, which have gained importance over the last decades, given their higher operation flexibility. Batch processes typically generate data structures with the following three components: a table with the initial conditions of each batch (batch recipe, charge conditions), **Z**; a three-way table of process operating conditions across time at each batch, for all batches, with dimensions [*batch run* × *variable* × *time*], **X**; and another table of product quality measurements, **Y**. Techniques such as multi-way PCA and PLS were developed in order to accommodate this type of structures for process monitoring and prediction proposes (Nomikos & MacGregor, 1995; Westerhuis *et al.*, 1999).

When the number of variables becomes very large, the monitoring and diagnosis procedures can become quite cumbersome and difficult to interpret. Under these conditions, if the variables have some natural grouping, like belonging to different production sections or product streams, the analysis can be carried out by retaining this natural blocking in order to make the interpretation of results easier. For that purpose, one may apply multiblock and hierarchical PLS or PCA techniques (MacGregor *et al.*, 1994; Smilde *et al.*, 2003; Westerhuis *et al.*, 1998).

3.3.2 Image Analysis

With the growing availability of inexpensive digital imaging systems (Bharati & MacGregor, 1998), new approaches were developed to take advantage of the information provided by this particular type of sensors. For instance, one can monitor the performance of an industrial boiler by taking successive digital images (RGB) of the turbulent flame, and using them to access operation status (Yu & MacGregor, 2004a; Yu & MacGregor, 2004b). Even though the images change rapidly, their projection onto the latent variable space (using PCA) is quite stable at a given operating condition. However, the projections do change significantly if the feed or operating conditions suffer modifications. The success of PCA in the extraction of a stable pattern for a given operating condition, is compatible with a view of the variation in the RGB intensities for all the pixels (usually forming a 256×256 array) as a latent variable process of the type (3.4), whose number of latent variables is indeed quite low (the monitoring scheme is essentially based on the first two latent variables). This kind of approach is also being used for on-line product quality control in the food industry and others where visual inspection plays a key role.

3.3.3 Multivariate Calibration

In multivariate calibration, one aims to build a model based on multivariate spectral data (**X** block) and concentrations of the solutions used to produce such spectra (**Y** block), in order to predict what the concentration of the specimens (analytes) will be when new samples become available, based only on quick measurements made by a spectrometer, and thus avoiding lengthy laboratory analytical procedures. The fact that spectrometer data usually follow Beer's law (the resulting spectrum is a linear combination of the pure component chemical spectra, appropriately weighted by their composition), provides a strong theoretical motivation for the use of model (3.5), and therefore, both PCR and PLS algorithms are extensively used in this context (Estienne *et al.*, 2001; Martens & Naes, 1989).

3.3.4 Soft Sensors

Soft sensors consist of inferential models that provide on-line estimates for the values of interesting properties, based on readily available measurements, such as temperatures, pressures and flow rates. This is particularly appealing in situations where the equipment required to measure those properties is expensive and difficult to implement on-line, but they can also be used in parallel, providing a redundant mechanism to monitor the measurement devices performance (Kresta *et al.*, 1994; MacGregor & Kourti, 1998).

3.3.5 Experimental Design

Experimental design procedures using latent variables, instead of the original ones, can reduce the number of experiments needed to cover the whole space of interest. This happens because, by moving together groups of variables in the latent variable modelling frameworks, the effective number of independent variables becomes greatly reduced, while the operation constraints that motivate such groupings are implicitly taken into account (Gabrielsson *et al.*, 2002; Wold *et al.*, 1986).

3.3.6 Quantitative Structure Activity Relationships (QSAR)

The goal in this field is to relate the structure and physico-chemical properties of the compounds (**X**-block) with their macroscopic functional, biologic or pharmaceutical properties, such as carcinogenicity, toxicity, degradability, response to treatment, among others (**Y**-block). The **X**-block variables may be melting points, densities or

parameters derived from the underlying molecular structures. Therefore, the goal here is to build simple models relating the two groups of variables for prediction of the biologic activity or pharmaceutical properties for a wider set of compounds (whose structure and physico-chemical properties are known, i.e., the **X**-block properties) from the knowledge of a limited number of fully characterized representatives (where both blocks of variables are known), or to optimize the structure in order to improve, in some sense, the activity variables. For instance, it may be required to predict the performance of a drug candidate, or just to know which properties regulate the response of the **Y**-block variables, so that we can modify compounds or search for others that match the required goal (Eriksson *et al.*, 2001). This is another field where latent variable models, and in particular PLS, have found great success, given the presence of strong relationships among variables belonging to each block and between the two blocks (see also Burhnam *et al.*, 1999, and references therein).

3.3.7 Product Design, Model Inversion and Optimization

In this context, latent variable models estimated using historical data from a given process, where process constraints and operating policies are already implicitly incorporated in the data correlation structure, are used to address different tasks. In product design, the model is used to find an operating window where a product can be manufactured with a desired set of properties (Jaeckle & MacGregor, 2000; MacGregor & Kourti, 1998). Such operating windows are derived from a definition of the desired quality specifications for the new product and an inversion over the latent variable model, from the **Y** to the **X** space. The solution thus found will not only comply with these properties, but also be compatible with past operating policies (Jaeckle & MacGregor, 1998). Such a model can also be used for optimization purposes, in particular to find the "best" operating conditions (Yacoub & MacGregor, 2004) and the "best" policies for batch process control (Flores-Cerrilo & MacGregor, 2004).

3.4 Wavelet Theory

3.4.1 Brief Historical Note

Although distinct roots regarding research lines linked to wavelets can be traced back to earlier times, the starting point for their modern developments is usually set in the early

80's, under the efforts pursued by the French geophysicist Jean Morlet, in connection with his work on the analysis of echo signals (direct reflections or backscattering) in oil prospecting (Hubbard, 1998). Morlet's empirical approach was brought to the attention of theoretical physicist Alex Grossman, who began collaborating with Morlet in the interpretation of the good results obtained, publishing the first paper where the word "wavelet" appears (Soares, 1997). The French mathematician Yves Meyer quickly acknowledged the connections between Morlet and Grossman's developments and a classic result from harmonic analysis (Calderón identity), getting into the field also in collaboration with the French mathematician Stéphan Mallat, who later on established the connection between the pure mathematical role of wavelets (*wavelets series expansions*) and a class of algorithms already developed in some applied fields, under different names, such as *multiresolution signal processing* from computer vision, *pyramid algorithms* from image processing, *subband coding* and *filter banks* from signal processing and *quadrature mirror filters* from digital speech processing (Bruce *et al.*, 1996; Burrus *et al.*, 1998; Hubbard, 1998; Rioul & Vetterli, 1991), giving mathematical depth to all of them, while providing, at the same time, strong and intuitive concepts, such as the notion of approximations and details as projections to particular subspaces of $L^2(\mathbb{R})$. Important contributions were also made in this context by the Belgian mathematical physicist Ingrid Daubechies, namely in developing a family of orthogonal wavelet transforms with compact support, that found wide application in many different fields, strongly contributing to the boost of activity in the development of multiscale approaches in connection with Mallat's multiresolution decomposition analysis framework. More historical details can be found elsewhere (Hubbard, 1998; Meyer & Ryan, 1993; Soares, 1997).

The list of books dedicated to the subject of wavelet theory is already quite extensive, ranging from basic level introductions (Aboufadel & Schlicker, 1999; Burrus *et al.*, 1998; Chan, 1995; Hubbard, 1998; Walker, 1999), encompassing more thorough treatments (Mallat, 1998; Strang & Nguyen, 1997), texts that follow a more mathematical-oriented approach (Chui, 1992; Kaiser, 1994; Walter, 1994), applications (Chau *et al.*, 2004; Cohen & Ryan, 1995; Motard & Joseph, 1994; Percival & Walden, 2000; Starck *et al.*, 1998; Vetterli & Kovačević, 1995), and more technically advanced presentations (Daubechies, 1992), as well as review articles (Alsberg *et al.*, 1997; Rioul

& Vetterli, 1991). Reference is made to these sources for more elaborate discussions (see also Reis, 2000 and Soares, 1997, for introductory texts written in Portuguese).

In the next subsections a brief presentation is provided for the purposes of following later sections of this thesis, mainly centred on orthogonal wavelet transforms[17] and practical implementation issues, along with some motivations regarding the success of applying wavelets in data analysis tasks.

3.4.2 Motivation

Data acquired from natural phenomena, economic activities or industrial plants, usually do present complex patterns with features appearing at different *locations* and with different *localizations* either in time or frequency (Bakshi, 1999). To illustrate this point, let us consider Figure 3.3, where an artificial signal is presented, composed by superimposing several deterministic and stochastic features, each one with its own characteristic time/frequency pattern. The signal deterministic features consist of a ramp, that begins right from the start, a step perturbation at sample 513, a permanent oscillatory component, and a spike at observation number 256. The stochastic feature consist of additive Gaussian white noise, whose variance increases after sample number 768. Clearly these events have different time/frequency locations and localizations: for instance, the spike is completely localized in the time axis, but fully delocalized in the frequency domain; on the other hand, the sinusoidal component is very well localized in the frequency domain but spreads over the whole time axis. White noise contains contributions from all the frequencies and its energy is uniformly distributed in the time/frequency plane, but the linear trend is essentially a low frequency perturbation, and its energy is almost entirely concentrated in the lower frequency bands. All these patterns appear simultaneously in the signal, and, therefore, one should be able to deal with them, without compromising one kind of features over the others. This can only be done, however, if we adopt the suitable "mathematical language" for efficiently describing data with such multiscale characteristics.

[17] For information regarding topics involving other types of wavelet basis, such as biorthogonal basis and overcomplete expansions, reference is made to the relevant literature (Daubechies, 1992; Kaiser, 1994; Mallat, 1998; Vetterli & Kovačević, 1995).

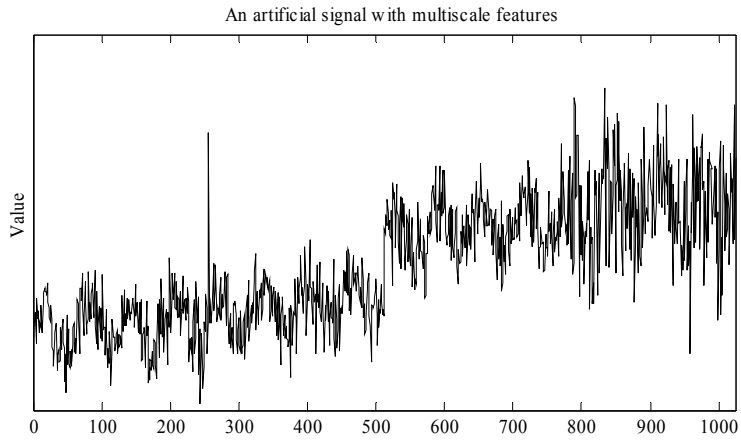


Figure 3.3. An artificial signal containing multiscale features, which results from the sum of a linear trend, a sinusoid, a step perturbation, a spike (deterministic features with different frequency localization characteristics) and white noise (a stochastic feature whose energy is uniformly distributed in the time/frequency plane).

Transforms, like the Fourier transform, provide alternative ways of representing raw data (i.e., playing the role of alternative "mathematical languages"), as an expansion of basis functions multiplied by the transform coefficients. These coefficients constitute the "transform", and, if the methodology is properly chosen, data analysis becomes much more efficient and effective when it is conducted over them, instead of over the original raw data. For instance, Fourier transform is the adequate mathematical "language" for describing periodic phenomena or smooth signals, since the nature of its basis functions allows for compact representations of such trends, meaning that only a few coefficients are needed in order to provide a good representation of the signal. The same applies, in other contexts, to other classical single-scale linear transforms (Bakshi, 1999; Kaiser, 1994; Mallat, 1998), such as the one based on the discrete Dirac-δ function or the windowed Fourier transform. However, none of these single-scale linear transforms are able to cope effectively with the diversity of features present in signals such as the one illustrated in Figure 3.3. A proper analysis of this signal, using these techniques, would require a large number of coefficients, indicating that they are not "adequate" languages for a compact translation of its key features in the transform domain. This happens because the form of the time/frequency windows (Mallat, 1998; Vetterli & Kovačević, 1995), associated with their basis functions (Figure 3.4), does not

change across the time/frequency plane, in order to effectively cover the localized high energy zones of the several features present in the signal.

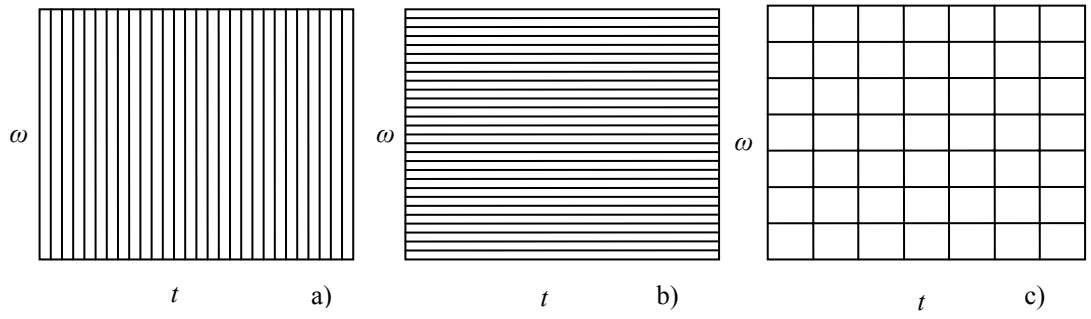


Figure 3.4. Schematic illustration of the time/frequency windows associated with the basis function for the following linear transforms: (a) Dirac-δ transform, (b) Fourier transform and (c) windowed Fourier transform.

Therefore, in order to cope with such multiscale features, a more flexible tiling of the time/frequency space is required, which can be found by adopting wavelets as basis functions (Figure 3.5), whose expansion coefficients are called wavelet transform. In practice, it is often the case that signals are composed of short duration events of high frequency and low frequency events of long duration. This is exactly the kind of tilling that a wavelet basis does provide, since the relative frequency bandwidth of these basis functions is a constant (i.e., the ratio between a measure of the size of the frequency band and the mean frequency,[18] $\Delta\omega/\omega$, is constant for each wavelet function), a property also referred to as a "constant-Q" scheme (Rioul & Vetterli, 1991).

[18] The location and localization of the time and frequency bands, for a given basis function, can be calculated from the first moment (the mean, a measure of location) and the second centred moment (the standard deviation, a measure of localization), of the basis function and its Fourier transform. The localization measures define the form of the boxes that tiles the time/frequency plane in Figure 3.4 and Figure 3.5. However, the time and frequency widths (i.e., localization) of these boxes do always conform to the lower bound provided by the Heisenberg principle ($\sigma(g)\cdot\sigma(\hat{g}) \geq 1/2$, where $\hat{g}$ represents the Fourier transform of *g*; Kaiser, 1994; Mallat, 1998). These boxes are often referred to as "Heisenberg boxes".

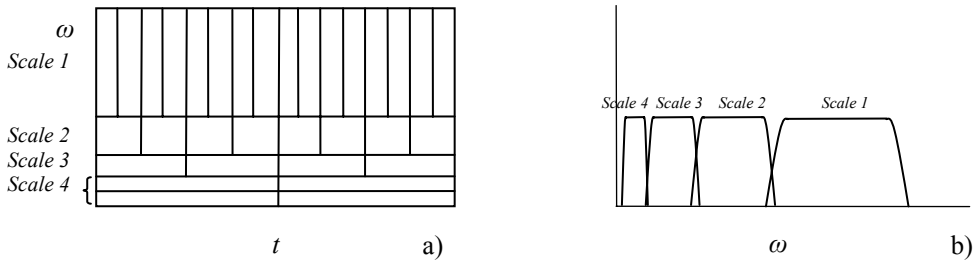


Figure 3.5. Schematic representation of the tiling of the time-frequency plane provided by the wavelet basis functions (a), and an illustration of how wavelets divide the frequency domain (b), where we can see that they work as bandpass filters. The shape of the windows and frequency bands, for a given wavelet function, depend upon the scale index value: for low values of the scale index, the windows have good time localizations and cover a long frequency band; windows with high values of the scale index have large time coverage with good frequency localization.

Wavelets are a particular type of functions whose location and localization characteristics in time/frequency are ruled by two parameters: both the localization in this plane and location in the frequency domain are determined by the scale parameter, *s*; the location in the time domain is controlled by the time translation parameter, *b*. Each wavelet, $\psi_{s,b}(t)$, can be obtained from the so called "mother wavelet", $\psi(t)$, through a scaling operation (that "stretches" or "compresses" the original function, establishing its form), and a translation operation (that controls its positioning in the time axis):

$$\psi_{s,b}(t) = \frac{1}{\sqrt{|s|}}\psi\left(\frac{t-b}{s}\right) \tag{3.13}$$

The shape of the mother wavelet is such that it does have an equal area above and below the time axis, which means that, besides having a compact localization in this axis, they should also oscillate around it, features from which derives the name of "wavelets" (small waves). In the Continuous Wavelet Transform (CWT), scale and translation

parameters can vary continuously, leading to a redundant transform[19] (a 1D signal is being mapped onto a 2D function). Therefore, in order to construct a basis set, it is sometimes possible to sample them appropriately, so that the set of wavelet functions parameterized by the new indices (scale index, *j*, and translation or shift index, *k*) covers the time-frequency plane in a non-redundant way. This sampling consists of applying a dyadic grid in which *b* is sampled more frequently for lower values of *s*, and *s* grows exponentially with the power of 2:

$$\psi_{j,k}(t) = \psi_{s,b}(t)\Big|_{\substack{s=2^j \\ b=k\cdot 2^j}} = \frac{1}{2^{j/2}}\psi\left(\frac{t-k\cdot 2^j}{2^j}\right) = \frac{1}{2^{j/2}}\psi\left(\frac{t}{2^j}-k\right) \tag{3.14}$$

The set of wavelet functions in (3.14) forms a basis for the space of all square integrable functions (Kreyszig, 1978), $L^2(\mathbb{R})$, which are infinite dimensional entities (functions). However, in data analysis, we almost always deal with vectors and matrices (data tables, images), which are dimensionally finite, but we still can use the above concepts with finite dimension entities, as explained in Section 3.4.4.

3.4.3 Multiresolution Decomposition Analysis

Working in a hierarchical framework for consistently representing images with different levels of resolution, i.e., containing different amounts of information regarding what is being portrayed, Stephane Mallat developed the unifying concept of *Multiresolution Approximation* (Mallat, 1989, 1998). A Multiresolution Approximation is a sequence, $\{V_j\}_{j\in\mathbb{Z}}$, of closed subspaces of $L^2(\mathbb{R})$, with the following 6 properties:

1. $\forall(j,k)\in\mathbb{Z}^2, f(t)\in V_j \Leftrightarrow f(t-2^j k)\in V_j;$ (3.15)

2. $\forall j\in\mathbb{Z}, V_{j+1}\subset V_j;$ (3.16)

[19] The redundancy of CWT is not necessarily undesirable, as it is translation-invariant (a property that is not shared by orthogonal wavelet transforms) and the coefficients do not have to be calculated very precisely in order to still obtain good reconstructions (Daubechies, 1992; Hubbard, 1998).

3. $\forall j \in \mathbb{Z}, f(t) \in V_j \Leftrightarrow f(\frac{t}{2}) \in V_{j+1}$; (3.17)

4. $\lim_{j \to +\infty} V_j = \bigcap_{j=-\infty}^{+\infty} V_j = \{0\}$; (3.18)

5. $\lim_{j \to -\infty} V_j = Closure \bigcup_{j=-\infty}^{+\infty} V_j = L^2(\mathbb{R})$; (3.19)

6. There exists ϕ such that $\{\phi(t-k)\}_{k \in \mathbb{Z}}$ is a Riesz basis of V_0. (3.20)

The first property states that any translation applied to a function belonging to the subspace V_j, proportional to its scale (2^j), generates another function still belonging to the same subspace. The second one refers that any entity in V_{j+1} also belongs to V_j, i.e., $\{V_j\}_{j \in \mathbb{Z}}$ is a sequence of nested subspaces: $\cdots \subset V_{j+1} \subset V_j \subset V_{j-1} \subset \cdots \subset L^2(\mathbb{R})$. In practice this means that projections to approximation functions with higher scale indices should originate coarser versions of the original function (or a lower resolution, coarser version of the original image), whereas projections to the richer approximation spaces, with lower scale indices, should result in finer versions of the projected function (or a finer version of the original image, i.e., with higher resolution). Property 3 requires that any dilation ("stretching") by a factor of two, applied to a function belonging to subspace V_j, results in a function belonging to the next coarser subspace V_{j+1}. However, if we keep "stretching" it, in the limit, when $j \to +\infty$, this function becomes a constant. This means that, in order for this limiting case still belong to $L^2(\mathbb{R})$, it must coincide with the constant zero function, $\{0\}$. This is the only function that belongs to all the approximation spaces, from the finest (lower scale indices) to the coarsest (higher scale indices), as stated by property 4. We can also conclude from this property that the projection to coarser approximation spaces successively originates both coarser and *residual* approximations of the original functions:

$$\lim_{j \to +\infty} \Pr_{V_j} f = 0 \ in \ L^2(\mathbb{R}) \Leftrightarrow \lim_{j \to +\infty} \left\| \Pr_{V_j} f \right\| = 0 \quad (3.21)$$

On the other hand, following from property 5, any function in $L^2(\mathbb{R})$ can successively be better approximated by a sequence of projections onto increasingly finer subspaces, i.e.:

$$\lim_{j\to-\infty} \mathrm{Pr}_{V_j} f = f \ in L^2(R) \Leftrightarrow \lim_{j\to-\infty} \left\| f - \mathrm{Pr}_{V_j} f \right\| = 0 \tag{3.22}$$

The last property concerns the existence of a Riesz basis for the space V_0, that consists of the so called *scaling function*, $\phi(t)$, along with its integer translations. In what follows this basis is an orthonormal one, which, according to properties 1 and 3, means that the set $\{\phi_{j,k}\} = \left\{ 2^{-\frac{j}{2}} \phi\left(2^{-j}t - k\right) \right\}$ is an orthonormal basis for V_j. Therefore, we have at this point a well characterized sequence of nested subspaces, with basis functions that result from translation/scaling operations applied to the scaling function.

Let us now introduce a complementary concept to the approximation subspaces: the *detail subspaces*, $\{W_j\}_{j\in\mathbb{Z}}$. As V_{j+1} is a proper subspace of V_j ($V_{j+1} \subset V_j, V_{j+1} \neq V_j$), we will call to the orthogonal complement of V_{j+1} in V_j, W_{j+1}. Therefore, we can write $V_j = V_{j+1} \overset{\perp}{\oplus} W_{j+1}$, which means that any function in V_j can univocally be given by a sum of elements belonging to the approximation space V_{j+1} and to the detail space W_{j+1}. These elements are just the projections onto these subspaces. As $V_j \subset V_{j-1}$, we can also state that $V_{j-1} = V_j \overset{\perp}{\oplus} W_j = V_{j+1} \overset{\perp}{\oplus} W_{j+1} \overset{\perp}{\oplus} W_j$. This means that, if we have a function, a signal or an image belonging to V_0, f_0, we can represent it as a projection into the approximation level at scale j, f_j, plus all the details relative to the scales in between ($\{w_i\}_{i=1,\ldots,j}$), since

$$V_0 = V_j \overset{\perp}{\oplus} W_j \overset{\perp}{\oplus} \cdots \overset{\perp}{\oplus} W_2 \overset{\perp}{\oplus} W_1 \tag{3.23}$$

In terms of projection operations:

$$f_0 = f_j + \sum_{i=1}^{j} w_i \Leftrightarrow f_0 = \mathrm{Pr}_{V_j} f_0 + \sum_{i=j}^{1} \mathrm{Pr}_{W_i} f_0 \tag{3.24}$$

It can be shown that an orthonormal basis for the details space W_j can be given by the set of wavelet functions, $\{\psi_{j,k}\}_{k\in\mathbb{Z}}$. These basis sets (for different scales) are mutually orthogonal, as they span orthogonal subspaces of $L^2(\mathbb{R})$. By extending decomposition (3.23) in order to incorporate all the scales, and considering properties 4 and 5, we can

conclude that $L^2(\mathbb{R}) = \overset{\perp}{\bigoplus_{i=-\infty}^{+\infty}} W_i$, meaning that the wavelets functions with the discrete parameterization do indeed form a basis of this space. The projections, f_j and $\{w_i\}_{i=1,\ldots,j}$ in (3.24), can adequately be written in terms of the linear combination of basis functions (3.25) multiplied by the expansion coefficients, calculated as inner products of the signal and basis functions (3.26):

- *Approximation coefficients:* $a_k^j \ (k \in \mathbb{Z})$;
- *Details coefficients:* $d_k^i \ (i = 1, \ldots, j;\ k \in \mathbb{Z})$.

These are usually referred to as the (discrete) *wavelet transform* or *wavelet coefficients*:

$$f_0 = \sum_k a_k^j \phi_{j,k} + \sum_{i=1}^{j} \sum_k d_k^i \psi_{i,k} \tag{3.25}$$

where

$$a_k^j = \left\langle f_j, \phi_{j,k} \right\rangle, \ d_k^i = \left\langle f_j, \psi_{i,k} \right\rangle \tag{3.26}$$

Still within the scope of the multiresolution approximation framework, Mallat (1989) proposed a very efficient recursive scheme for the computation of wavelet coefficients, equations (3.27) and (3.28), as well as for signal reconstruction, equation (3.29), that basically consists of implementing a pyramidal algorithm, based upon convolution with quadrature mirror filters, a well known technique in the engineering discrete signal processing community:

- *Signal analysis or decomposition*

$$a_k^{j+1} = \sum_n h_{n-2k} \cdot a_n^j \tag{3.27}$$

$$d_k^{j+1} = \sum_n g_{n-2k} \cdot a_n^j \tag{3.28}$$

- *Signal synthesis or reconstruction*

$$a_k^j = \sum_n h_{k-2n} \cdot a_n^{j+1} + \sum_n g_{k-2n} \cdot d_n^{j+1} \tag{3.29}$$

where $\{h_i\}_{i\in\mathbb{Z}}$ and $\{g_i\}_{i\in\mathbb{Z}}$ are the low-pass and high-pass filter coefficients, respectively, whose values are intimately connected (Aboufadel & Schlicker, 1999; Daubechies, 1992; Mallat, 1998; Strang & Nguyen, 1997).

The recursive nature of the computation scheme underlying equations (3.27)-(3.29) is illustrated in Figure 3.6 (for the analysis or decomposition algorithm) and Figure 3.7 (for the synthesis or reconstruction algorithm).

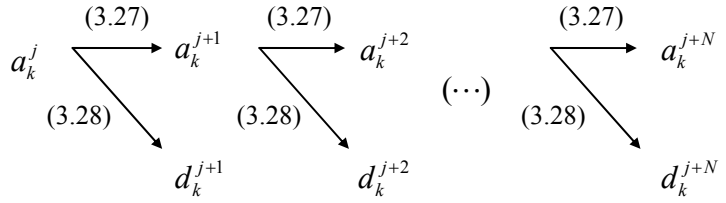


Figure 3.6. Schematic representation of recursive scheme for the computation of wavelet coefficients (analysis algorithm). It is equivalent to performing convolution with an analysis filter followed by dyadic downsampling.

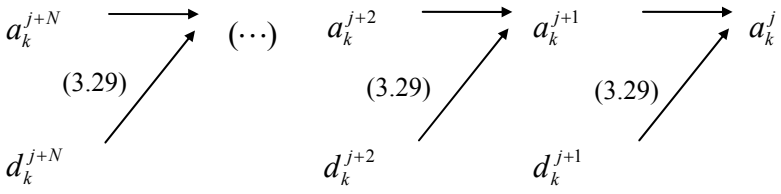


Figure 3.7. Schematic representation of recursive scheme for reconstruction of the signal from the wavelet coefficients (synthesis algorithm). Each stage consists of an upsampling operation followed by convolution with the synthesis filter and adding of outputs.

The above operations can also be formulated in matrix terms, where the analysis and synthesis procedures are the result of applying certain transformation matrices to raw data or wavelet coefficients, respectively (Bakshi, 1998; Reis, 2000; Yoon & MacGregor, 2004). For instance, the analysis process can be represented by:

$$Y = W_A X \tag{3.30}$$

where

$$\overbrace{\begin{bmatrix} \mathbf{a}_J \\ \mathbf{d}_J \\ \vdots \\ \mathbf{d}_1 \end{bmatrix}}^{Y} = \overbrace{\begin{bmatrix} \mathbf{H}_J \\ \mathbf{G}_J \\ \vdots \\ \mathbf{G}_1 \end{bmatrix}}^{W_A} \overbrace{\begin{bmatrix} \mathbf{a}_0 \end{bmatrix}}^{X} \tag{3.31}$$

with $\mathbf{a}_j$ and $\mathbf{d}_j$ $\left(n/2^j \times 1\right)$ being vectors of wavelet coefficients (j is the scale index, $j = 1:J$), $\mathbf{H}_j$ and $\mathbf{G}_j$ $\left(n/2^j \times n\right)$ matrices of coefficients entirely defined by the wavelet filter coefficients, $\{h_i\}_{i\in\mathbb{Z}}$ and $\{g_i\}_{i\in\mathbb{Z}}$, and $\mathbf{a}_0$ the original raw signal, under the form of a $\left(n \times 1\right)$ vector.

For an orthogonal wavelet transform, the analysis matrix, W_A, is orthogonal (unitary, in the more general, complex case), which means that the synthesis matrix, W_S, which is such that

$$f = W_S Y \tag{3.32}$$

can be simply defined by the transpose (real case): $W_S = W_A^T$ (in the complex case, the hermitian transpose should be used instead, i.e., $W_S = \overline{W}_A^T$, where $\overline{W}$ is the complex conjugate of W; the presentation that follows is centred around the real case, as this is the one encountered in most of the applications found in Chemical Engineering). Thus,

$$W_S = \begin{bmatrix} \mathbf{H}_J^T & \mathbf{G}_J^T & \cdots & \mathbf{G}_1^T \end{bmatrix} \tag{3.33}$$

from which follows that

$$\begin{aligned} X = W_S W_A X = \begin{bmatrix} \mathbf{H}_J^T & \mathbf{G}_J^T & \cdots & \mathbf{G}_1^T \end{bmatrix} \begin{bmatrix} \mathbf{H}_J \\ \mathbf{G}_J \\ \vdots \\ \mathbf{G}_1 \end{bmatrix} = \\ = \underbrace{\mathbf{H}_J^T \mathbf{H}_J X}_{f_J} + \underbrace{\mathbf{G}_J^T \mathbf{G}_J X}_{w_J} + \cdots + \underbrace{\mathbf{G}_1^T \mathbf{G}_1 X}_{w_1} \end{aligned} \tag{3.34}$$

where f_J, w_J and w_1 are $(n \times 1)$ vectors, representing the contribution for X arising from the projection to the approximation space V_J and from the projections to the detail spaces at different scales $\{W_j\}_{j=1:J}$.

As an illustration, we can decompose the signal in Figure 3.3, that contains 2^{10} points at scale $j=0$, into a coarser, lower resolution version at scale $j=5$ with 2^5 approximation coefficients appearing in the expansion ($f_5 = \sum_{k=0}^{2^5-1} a_k^5 \phi_{5,k}$) plus all the detail signals from scale $j=1$ (with 2^9 detail coefficients, $w_1 = \sum_{k=0}^{2^9-1} d_k^1 \psi_{1,k}$) until scale $j=5$ (with 2^5 detail coefficients, $w_5 = \sum_{k=0}^{2^5-1} d_k^1 \psi_{5,k}$). The total number of wavelet coefficients is equal to the cardinality of the original signal, thus no information is "created" or "disregarded", but simply transformed ($2^{10} = 2^5 + 2^5 + 2^6 + \cdots + 2^9$). The projections onto the approximation and detail spaces are presented in Figure 13.7, where we can see that the deterministic and stochastic features appear quite clearly separated, according to their time/frequency location and localization: coarser deterministic features (ramp and step perturbation) appear in the coarser version of the signal (containing the lower frequency contributions), the sinusoid is captured in the detail at scale $j=5$, noise features appear quite clearly at high frequency bands (details for $j=1,2$), where the increase of variance is noticeable, as well as the spike at observation 256 (another high frequency perturbation.) This illustrates the ability of wavelet transforms to separate deterministic and stochastic contributions present in a signal, according to their time/frequency locations.

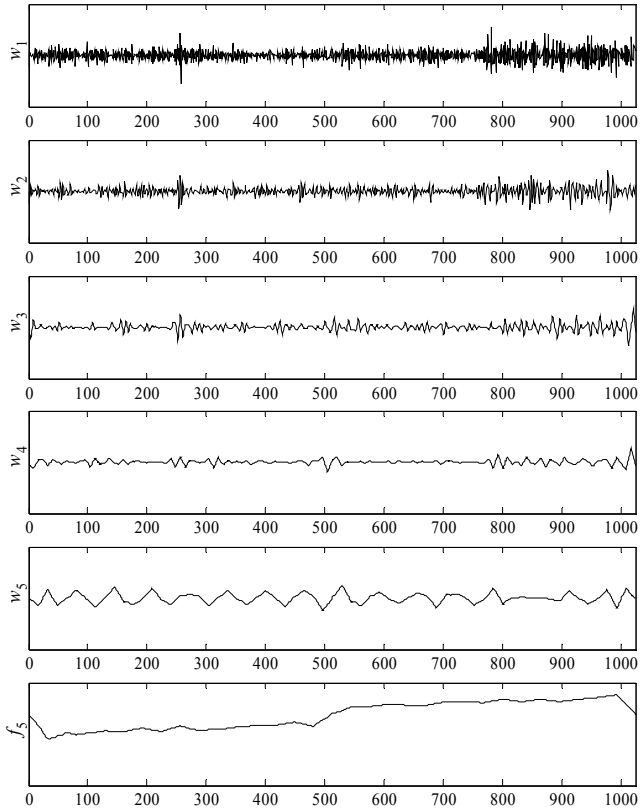


Figure 3.8. The signal in Figure 3.3 decomposed into its coarser version at scale $j = 5$ plus all the details lost across the scales ranging from $j = 1$ to $j = 5$. The filter used here is the Daubechies's compactly supported filter with 3 vanishing moments.[20]

3.4.4 Practical Issues on the Use of Wavelet Transforms

In practice, for finite dimensional elements (data arrays), it is usually assumed that the available data is already the projection onto space V_0 (Bakshi, 1998), f_0, and the computation of the wavelet coefficients proceeds through Mallat's efficient recursive

[20] A wavelet has *p* vanishing moments if $\int_{-\infty}^{+\infty} t^k \psi(t) dt = 0$ for $0 \le k < p$. This is an important property in the fields of signal and image compression, since it can induce a higher number of low magnitude detail coefficients, if the signal does have local regularity characteristics.

analysis algorithm given by equations (3.27)-(3.28). More elaborate initialization strategies are discussed elsewhere (Daubechies, 1992; Mallat, 1998; Strang & Nguyen, 1997). Therefore, we essentially apply the analysis and reconstruction quadrature mirror filters associated to a given wavelet, without using any wavelet function explicitly. In fact, very often wavelets do not even have a closed formula in the time domain, even though they can be plotted as accurately as required, by iterating over such filters (Strang & Nguyen, 1997).

When transforming finite length signals using filters other than the Haar filter or a family of boundary corrected wavelet filters (Depczynsky *et al.*, 1999; Mallat, 1998), one has to deal with the boundary problem issue, derived from the lack of data for applying the filters near the signal boundaries. Therefore, the signal should be somehow expanded, and several strategies are available for doing such, as for instance: "zero-padding" (extend by adding zeros), "wraparound" (extend by periodicity), symmetric extension (extend by reflection) (Strang & Nguyen, 1997), linear padding (Trygg *et al.*, 2001; Trygg & Wold, 1998) and level padding (Teppola & Minkkinen, 2000). Trygg *et al.* (2001) proposed a different approach, that does not require extending the original signal (with the subsequent increase in computation load), called the "set-aside" approach, which consists of setting aside the last low-pass coefficient, adding it to the details coefficients vector of that scale, whenever the signal at this scale is not of even length. This strategy allows the computations to pursue to higher scales, as the vector with the remaining low-pass filters has now even length.

Another issue always present is the choice of the wavelet filter. Teppola & Minkkinen (2001) used the Symmlet-10 wavelet in their work, and referred, as a rule of thumb, that "*smooth wavelet functions should be used with smooth data*". Trygg *et al.* (2001) referred another rule of thumb, according to which one should "*selected a wavelet with more vanishing moments than twice the polynomial order of the interesting signal to analyze*". Staszewski (1998) points out that the selection procedure is usually a trade-off between smoothness (differentiability) and compact support of the wavelet. In general terms, the author refers that " *(...) more compactly supported, and therefore less smooth wavelet functions, are better for non-stationary data with discontinuities, impulses or transients. (...) Less compactly supported, and therefore more smooth wavelet functions, are better for stationary, regular data or in cases where low level of compression error is required. (...) For regular, smooth, stationary data, more*

vanishing moments lead to smaller wavelet coefficients. However, for non-stationary, [irregular] data more vanishing moments lead to more large wavelet coefficients."

Trygg & Wold (1998) used the Daubechies wavelet with four vanishing moments because of its relatively short filter length (which means less computational load, as the overall number of calculations is roughly proportional to $2 \cdot C \cdot K$, where C is the number of coefficients in the filter and K the signal length; for more details on the quantification of computational load see Vogt & Tacke, 2001), and because other smoother wavelets, like the Symmlet-8 (Symmlet with eight vanishing moments), provided similar results. Teppola & Minkkinen (2000), on the other hand, used Symmlets-10 instead of wavelets from the Daubechies family, because the former are more symmetric, enabling better interpretation of the resulting coefficients. Alsberg *et al.* (1998) used the Symmlet-8, as it has a suitable shape for the kind of peaks founded on the spectra under analysis (infrared spectra), considering as "unsuitable" those wavelet forms that require more scales to be included in the reconstruction in order to achieve a similar performance.

Alsberg (2000) applied an optimal compression-related criterion to the mean spectrum of the data set, and identified the Symmlet-9 wavelet as the one that resulted in best performance. Pasti *et al.* (1999) also presented a systematic approach for selecting both the best wavelet filter and depth of decomposition for signal de-noising in the wavelet domain, using a cross-validation procedure.

The design of the wavelet filter can also be oriented towards the optimization of its predictive performance in wavelet regression applications (Coelho *et al.*, 2003; Galvão *et al.*, 2004).

Part IV-A

Single-Scale Data Analysis

So, a result without reliability (uncertainty) statement cannot be published or communicated because it is not (yet) a result. I am appealing to my colleagues of all analytical journals not to accept papers anymore which do not respect this simple logic.

P. de Bièvre , Accredit. Qual. Assur. (Editorial), **2** (1997) 269

Chapter 4. Generalized Multiresolution Decomposition Frameworks

Multiresolution decomposition (MRD) frameworks are instrumental when one needs to focus data analysis at a particular scale, or to separate the several contributions to the overall phenomenon, arising from different scales either in time or length. However, its implementation with real world industrial data does not constitute always a straightforward procedure, namely in situations where a fraction of data is missing (either at random, or when variables have different acquisition rates, i.e., multirate data). Furthermore, the wavelet-based MRD frameworks do not integrate explicitly measurement uncertainty information in their calculations, therefore leaving aside a piece of information that might be relevant for the posterior analysis goals, and that furthermore is becoming increasingly available for a wide range of measurement devices, following the recent developments on measurement methods and metrology, as well as the increasing enforcement driven by standardization organizations.[21]

[21] See e.g. resolution number 21 of CEN Technical Board in 2003, that resolves following the suggestions made by working group CEN/BT WG 122 "Uncertainty of Measurement", laid down in report BT N 6831.

Therefore, it is desirable to develop MRD strategies that could still be easily implemented under sparse data structures contexts, and that are able to integrate uncertainty information in order to make it available at each scale, allowing one to explore its potential in subsequent tasks. This research effort is also aligned with the current trend of "up-dating" classical data analysis approaches, formerly strictly based only upon raw data, to their uncertainty-based counterparts (Section 3.2).

In this chapter, we address the development of multiresolution decomposition methodologies that are able to cope with difficult data/uncertainty structures, often met in industrial practice, like missing data and heteroscedastic uncertainties. Furthermore, guidelines are provided regarding an adequate use of the proposed methodologies and several examples, from simulated situations to real world, industrial and laboratorial case studies, are used to illustrate their operation and practical utility under several application contexts, with rather different goals, such as *scale selection* and signal *de-noising*.

4.1 Uncertainty-Based MRD Frameworks

For the present purposes, a "multiresolution decomposition framework" is considered to be an algorithm developed in order to provide expansion coefficients of the type obtained with the wavelet decomposition procedure (see Section 3.4.3), that contain localized information in a certain region of the time/scale plane. For the classical situation, where no data is missing and uncertainty information is not explicitly considered, it reduces itself to the wavelet transform, where the basis functions of the expansion have well defined properties, established by design, and for which there is available a very efficient algorithm for computing the coefficients, as well as for reconstructing the signal back into the original domain (Mallat, 1989, 1998). However, this classical procedure can not be straightforwardly applied in less conventional situations, like those with missing data, something that occurs quite often in industrial scenarios, and, furthermore, does not explicitly take into account data uncertainties, when these are available.

Therefore, in the next subsections, three categories of methodologies are presented, devoted to situations with missing data and homoscedastic (constant) or heteroscedastic (varying) uncertainties (*Method 1* and *Method 2*), and to situations where there is no

missing data and uncertainties can be either homoscedastic or heteroscedastic (*Method 3*). These are referred to as "generalized" (Haar) MRD frameworks, as they reduce to this particular type of wavelet transform in the case of homoscedastic noise whose uncertainty is known *a priori*, without missing data, but are also able to handle the more complex situations where one or both of these complicating features do arise.[22]

4.1.1 Method 1: Adjusting Filter Weights According to Data Uncertainties

The Haar wavelet transform, perhaps the simplest and one of the most well known among the wavelet transforms, attributes a very clear meaning to its coefficients: approximation coefficients are averages over non-overlapping blocks of two successive elements, and detail coefficients correspond to the difference between this average and the first element of the block.[23] Cascading this procedure across the successive sets of approximation coefficients thus obtained, results in the Haar wavelet transform, which can be written simply as:

$$\begin{aligned} a_{\lceil k/2 \rceil}^{j+1} &= C_{\lceil k/2 \rceil} \cdot \left(a_k^j + a_{k+1}^j \right) \\ d_{\lceil k/2 \rceil}^{j+1} &= C_{\lceil k/2 \rceil} \cdot \left(a_{\lceil k/2 \rceil}^j - a_k^j \right) \end{aligned} \tag{4.1}$$

where

$$C_{\lceil k/2 \rceil} = \sqrt{2}/2 \tag{4.2}$$

with a_k^j and d_k^j being the approximation and detail coefficients relative to the scale indexed by j and shift indexed by k, respectively, and $\lceil x \rceil$ the smallest integer $n \geq x$. Such a computation procedure gives equal weight to both values participating in the calculation of the average (coarser approximation coefficient). However, in case there is uncertainty information available, regarding data under analysis, the averaging process

---

[22] The third methodology (*Method 3*), in fact, does not concern only the Haar transform, but *any* orthogonal wavelet transform.

[23] The averages and differences are scaled by a factor of $1/\sqrt{2}$, in order to preserve the signal's energy after transformation (Parseval relation; Kreyszig, 1978; Mallat, 1998).

can be modified in order to *increase the weight* given to the datum with *less uncertainty*, in the calculation of the coarser approximation coefficient. This can be achieved by using different and properly chosen averaging coefficients, to be applied to each datum, referred as $C_{\lceil k/2 \rceil}^{j+1,1}$ and $C_{\lceil k/2 \rceil}^{j+1,2}$, in order to reflect their varying nature along the scale and shift indices:

$$a_{\lceil k/2 \rceil}^{j+1} = C_{\lceil k/2 \rceil}^{j+1,1} \cdot a_k^j + C_{\lceil k/2 \rceil}^{j+1,2} \cdot a_{k+1}^j \tag{4.3}$$

Adequate weights can be set by adopting the MVUE (minimum variance unbiased estimator) equations for the (common) average, that define the following averaging coefficients, associated to each datum (Guimarães & Cabral, 1997; Montgomery & Runger, 1999):

$$C_{\lceil k/2 \rceil}^{j+1,1} = \frac{1 \Big/ u\left(a_k^j\right)^2}{1 \Big/ u\left(a_k^j\right)^2 + 1 \Big/ u\left(a_{k+1}^j\right)^2} \tag{4.4}$$

$$C_{\lceil k/2 \rceil}^{j+1,2} = 1 - C_{\lceil k/2 \rceil}^{j+1,1} \tag{4.5}$$

where $u(x)$ represents the uncertainty associated with x. Detail coefficients are computed through:

$$d_{\lceil k/2 \rceil}^{j+1} = C_{\lceil k/2 \rceil}^{j+1,1} \cdot \left(a_k^j - a_{\lceil k/2 \rceil}^{j+1}\right) = C_{\lceil k/2 \rceil}^{j+1,2} \cdot \left(a_{\lceil k/2 \rceil}^{j+1} - a_{k+1}^j\right) \tag{4.6}$$

where we can see that these coefficients are such that the equality preserves some resemblance relatively to the Haar case, namely regarding the terms inside brackets (i.e., the only difference in equation (4.6), with regard to the Haar case, relies on the varying coefficients and the scaling factor).

Data uncertainty associated with the approximation coefficients at scale *j* should also be propagated to the approximation and detail coefficients computed at the next coarser

scale, j+1, to allow for the specification of uncertainties associated with the coefficients computed at these scales, therefore enabling the averaging procedure to continue. This can be done by applying the general law of propagation of uncertainties to the present situation (ISO, 1993; Lira, 2002):

$$u\left(a_{\lceil k/2\rceil}^{j+1}\right)=\sqrt{\left(C_{\lceil k/2\rceil}^{j+1,1}\right)^2\cdot u\left(a_k^j\right)^2+\left(C_{\lceil k/2\rceil}^{j+1,2}\right)^2\cdot u\left(a_{k+1}^j\right)^2} \tag{4.7}$$

$$u\left(d_{\lceil k/2\rceil}^{j+1}\right)=\sqrt{\left(C_{\lceil k/2\rceil}^{j+1,1}\right)^2\cdot\left(u\left(a_k^j\right)^2+u\left(a_{k+1}^j\right)^2-2\cdot C_{\lceil k/2\rceil}^{j+1,1}\cdot u\left(a_k^j\right)^2\right)} \tag{4.8}$$

where it is assumed that errors affecting two successive observations are statistically independent from each other, although more complex error structures can also be considered under this framework. By conducting a multiresolution decomposition, using this procedure, more weight is given to the values with less associated uncertainties during the calculation of the approximation coefficients. In the limit, if a datum at scale j has a very high uncertainty associated with it, then this value will not contribute significantly to the calculation of the next approximation coefficient at scale j+1, and the correspondent detail coefficient will also have a very low magnitude, in agreement with the intuitive reasoning that, in fact, very little detail is lost in the replacement of the two values by their uncertainty-based weighted average (4.4), when one of them is not reliable at all.

Extending this reasoning even further, we can verify that this computation scheme offers an easy and coherent way to integrate missing data in the analysis, as a missing datum can be considered to be any finite number with an infinite uncertainty associated with it, which effectively removes it from equations (4.3)-(4.8). In this situation, the coarser approximation coefficient assumes the same value as the non-missing datum and the coarser detail coefficient is zero. Therefore, there is no need for any additional formal modification of *Method 1*, in order to accommodate for the presence of missing data.

When no missing data are present, and the uncertainties are homoscedastic, this multiresolution decomposition framework provides the same results as the Haar transform (up to a scaling factor of $2^{j/2}$, for the coefficients at scale j).

4.1.2 Method 2: Use Haar Wavelet Filter, Accommodate Missing Data and Propagate Data Uncertainties to Coarser Coefficients

In this second approach for incorporating data uncertainties in MRD, the averaging and differencing coefficients are kept constant and equal to the ones suggested by the Haar wavelet transform filters.

When there are no missing data, the uncertainties of the finer approximation coefficients are *propagated* to the coarser approximation and detail coefficients, using the law of propagation of uncertainties:

$$u\left(a_{\lceil k/2\rceil}^{j+1}\right)=u\left(d_{\lceil k/2\rceil}^{j+1}\right)=\sqrt{\left(\sqrt{2}/2\right)^2\cdot u\left(a_k^j\right)^2+\left(\sqrt{2}/2\right)^2\cdot u\left(a_{k+1}^j\right)^2} \tag{4.9}$$

where, as before, serial statistical independency is assumed for the errors. If there are missing data, we calculate the next coarser coefficients by successively applying the following rules to each new pair of approximation coefficients at scale j, $\left\{a_k^j, a_{k+1}^j\right\}$:

Table 4.1. Uncertainty-based MRD frameworks: table of rules for *Method 2*.

- **Rule 1.** No missing data $\Rightarrow$ use Haar and calculate uncertainties through (4.9);
- **Rule 2.** $\left\{a_k^j\right\}$ is missing $\Rightarrow \begin{cases} a_{\lceil k/2\rceil}^{j+1}=a_{k+1}^j, u\left(a_{\lceil k/2\rceil}^{j+1}\right)=u\left(a_{k+1}^j\right) \\ d_{\lceil k/2\rceil}^{j+1}=0, \quad u(d_{\lceil k/2\rceil}^{j+1})=0 \end{cases}$;
- **Rule 3.** $\left\{a_{k+1}^j\right\}$ is missing $\Rightarrow \begin{cases} a_{\lceil k/2\rceil}^{j+1}=a_k^j, u\left(a_{\lceil k/2\rceil}^{j+1}\right)=u\left(a_k^j\right) \\ d_{\lceil k/2\rceil}^{j+1}=0, u(d_{\lceil k/2\rceil}^{j+1})=0 \end{cases}$;
- **Rule 4.** $\left\{a_k^j, a_{k+1}^j\right\}$ are missing $\Rightarrow \begin{cases} a_{\lceil k/2\rceil}^{j+1}=\text{missing}, u\left(a_{\lceil k/2\rceil}^{j+1}\right)=\text{missing} \\ d_{\lceil k/2\rceil}^{j+1}=\text{missing}, u(d_{\lceil k/2\rceil}^{j+1})=\text{missing} \end{cases}$.

From the rules above we can see that, when there are no missing data, the procedure consists of applying the Haar wavelet with uncertainty propagation. But, when we have some missing data, it can also happen that it remains missing at coarser scales (Rule 4).

This fact is instrumental for analysing the information content at different scales, as described in the scale selection methodology referred further ahead (Section 4.4).

4.1.3 Method 3: Use Any Orthogonal Wavelet Filter and Propagate Data Uncertainties to Coarser Coefficients

Although noise is almost always present in industrial data sets, missing data is not always an issue. Therefore, for those situations where complete data sets are available, we would like to explore the benefits of using wavelet filter coefficients that were designed in some optimal sense, so that their good multiscale decomposition properties can be brought to the analysis. However, data uncertainties, if known, should also be incorporated as a way to allocate the available knowledge regarding raw data uncertainty to the approximation and detail coefficients computed. There is a situation where this task is particularly simple, that occurs when the uncertainties across the observations of each variable are homoscedastic and noise realizations are independent. In this case, it can be shown that all the approximation and detail coefficients at coarser scales have the same uncertainty as raw data at the finest scale (Jansen, 2001; Mallat, 1998). This can easily be checked by analysing equation (4.9) for the case of the Haar wavelet, but still holds for any other orthogonal wavelet family, as a consequence of the following theorem:

Theorem 4.1. For zero mean *i.i.d.* noise and orthogonal wavelet transforms (with the necessary boundary corrections), the covariance of noise affecting wavelet coefficients is the same as the covariance of the noise affecting raw data.

Proof. Consider the following $(n\times 1)$ vector of noisy observations, y, of the true signal, x, corrupted with noise, ε, both being also $(n\times 1)$ vectors:

$$y = x + \varepsilon \tag{4.10}$$

Applying the wavelet transformation corresponds to pre-multiplying these vectors by the transform matrix, W_A, leading to:

$$W_A y = W_A\left(x+\varepsilon\right) \Leftrightarrow \overbrace{W_A y}^{\tilde{y}} = \overbrace{W_A x}^{\tilde{x}} + \overbrace{W_A \varepsilon}^{\tilde{\varepsilon}} \Leftrightarrow \tilde{y} = \tilde{x} + \tilde{\varepsilon} \tag{4.11}$$

The covariance of the noise affecting the wavelet coefficients, $\tilde{\varepsilon}$, is given by:

$$\mathrm{cov}\left(\tilde{\varepsilon}\right)=\mathrm{cov}\left(W_A\varepsilon\right)=E\left\{\left(W_A\varepsilon\right)\left(W_A\varepsilon\right)^T\right\}=W_AE\left\{\varepsilon\varepsilon^T\right\}W_A^T=W_A\,\mathrm{cov}\left(\varepsilon\right)W_A^T \quad (4.12)$$

(note that $\mu_{\tilde{\varepsilon}}=W_A\mu_\varepsilon=W_A\mathbf{0}=\mathbf{0}$). As the noise is *i.i.d.*, $\mathrm{cov}\left(\varepsilon\right)=\sigma\mathbf{I}_n$, and noting that the wavelet transform matrix is unitary for orthogonal wavelet transforms, i.e., $W_AW_A^T=W_A^TW_A=\mathbf{I}_n$, it follows that:

$$\mathrm{cov}\left(\tilde{\varepsilon}\right)=\sigma W_AW_A^T=\sigma\mathbf{I}_n=\mathrm{cov}\left(\varepsilon\right) \quad (4.13)$$

■

For heteroscedastic situations, the law of propagation of uncertainties should be applied in order to calculate the uncertainties associated with the coefficients at coarser scales. When implemented with the Haar filter, this method coincides with *Method 2* for situations with no missing data, but, as opposed to *Method 2*, it also holds for other wavelet filters as well.

4.2 Guidelines on the Use of Generalized MRD Frameworks

Method 1, on one hand, and *Methods 2* and *3*, on the other, differ deeply on how they implement the incorporation of uncertainty information in their respective MRD frameworks. In this section we provide a general guideline about which type of approach to use and when. We introduce it through an illustrative example, which helps to clarify the underlying reasoning.

Let us consider an artificial, piecewise constant signal, where values are held constant in windows of $2^4=16$ successive values (Figure 4.1-*a*), to which proportional noise with uncertainties assumedly known is added. Using the noisy signal (Figure 4.1-*b*) it is possible to compute its approximations at coarser scales ($j=1,2,\ldots$), according to the two types of approaches (*Method 1* and *Methods 2-3*), and then to see which method performs better in the task of approximating the true signal when projected at the same scale, say *j*. The performance index used here is the mean square error between the approximation at scale *j*, calculated for the noisy signal and that for the true signal, MSE(*j*). Figure 4.1-*c* summarizes the results obtained for 100 of such simulations.

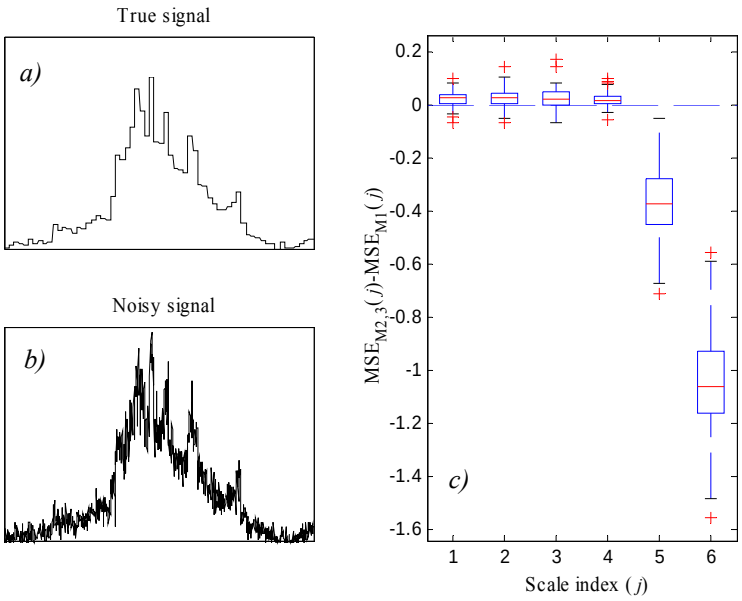


Figure 4.1. Illustrative example used for introducing a guideline regarding selection of the type of generalized MRD framework to adopt: (a) true signal used in the simulation; (b) a realization of the noisy signal and (c) box plots for the difference in MSE at each scale (j) obtained for the two types of methods, i.e. *Method 1* (M1) and *Methods 2-3* (M2,3), over 100 simulations.

These results illustrate a general guideline, according to which, from the strict point of view of the approximation ability at coarser scales, *Method 1* is more adequate then *Methods 2-3 for constant signals and for piecewise constant signals until we reach the scale where the true values begin to vary from (coarser) observation to (coarser) observation*, i.e., after which the piecewise constant behaviour stops. As the original signal has constant values along windows of 16 values, the piecewise constant pattern breaks down after scale $j = 4$.

This occurs because *Method 1* is based on the MVUE estimator of an underlying *constant* (or *common*) mean for two successive values, therefore leading to improved results when this assumption holds, at least approximately, as happens in the case of piecewise constant signals, being overtaken by the second type of methods (*Methods 2-3*) when such an assumption is no longer valid.

4.3 Uncertainty-Based De-Noising

In this section and the next one, two tasks where the generalized MRD frameworks can be used, with advantage over their classical counterparts, are addressed: uncertainty-based de-noising (this section) and "scale selection" (next section).

As already referred in Section 2.1, wavelets found great success in the task of "cleaning signals" from undesirable components of stochastic nature, often called in a general sense as "noise". If we are in such a position that we know the main noise features, namely measurement uncertainties, then we can use this additional piece of information to come up with simple but effective de-noising schemes. As an illustration, we will consider a smoothed version of a NIR spectrum as the "true" signal, to which heteroscedastic proportional noise was added. The standard de-noising procedure was then applied to the noisy signal, according to the following sequence of steps:

1. Decomposition of the signal into its wavelet coefficients;
2. Application of a thresholding technique to the calculated coefficients;
3. Reconstruction of the signal using the coefficients processed in stage 2.

This general procedure was tested with a classical implementation of the Haar wavelet transformation, using the threshold suggested by Donoho and Johnstone (1992), $T = \hat{\sigma}\sqrt{2\ln(N)}$, where $\hat{\sigma}$ is a robust estimator of noise (constant) standard deviation, along with a "Translation Invariant" extension of it, based on Coifman's "Cycle Spinning" concept (Coifman and Donoho, 1995):

"Average[Shift – De-noise – Unshift]"

where *all* possible shifts were used. We will call this alternative as "TI Haar".

These methods are to be compared with their counterpart procedures, that have the advantage of using available uncertainty information, referred as "Haar+uncertainty propagation" (i.e., *Methods 2* or *3*, because they coincide when there is no missing data), and "TI Haar+uncertainty propagation" (only *10 rotations* were used for this methodology).

For all of the alternatives we used the same wavelet (Haar), threshold constant ($\sqrt{2\ln(N)}$) and thresholding policy ("Hard Threshold"). Figure 4.2 presents the results obtained regarding MSE scores of the reconstructed signal (scale $j = 0$), relatively to

the true one, obtained after 100 realizations of additive noise. A clear improvement in MSE is found for the uncertainty-based methods. Figure 4.3 illustrates the de-noising effect for one of such realizations, where the more effective de-noising action provided by the uncertainty-based methods can be seen graphically. The smoothing action, due to the averaging scheme over several shifts, enhances the discontinuous nature of the de-noised signal obtained with the Haar wavelet filter.

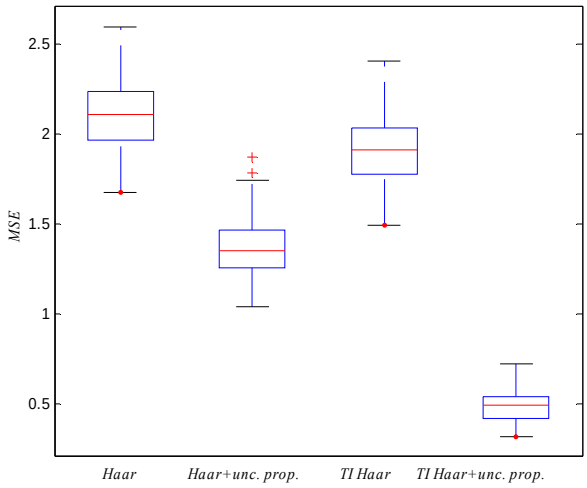


Figure 4.2. De-noising results associated with the four alternative methodologies ("Haar", "TI Haar", "Haar+uncertainty propagation" and "TI Haar+uncertainty propagation"), for 100 noise realizations.

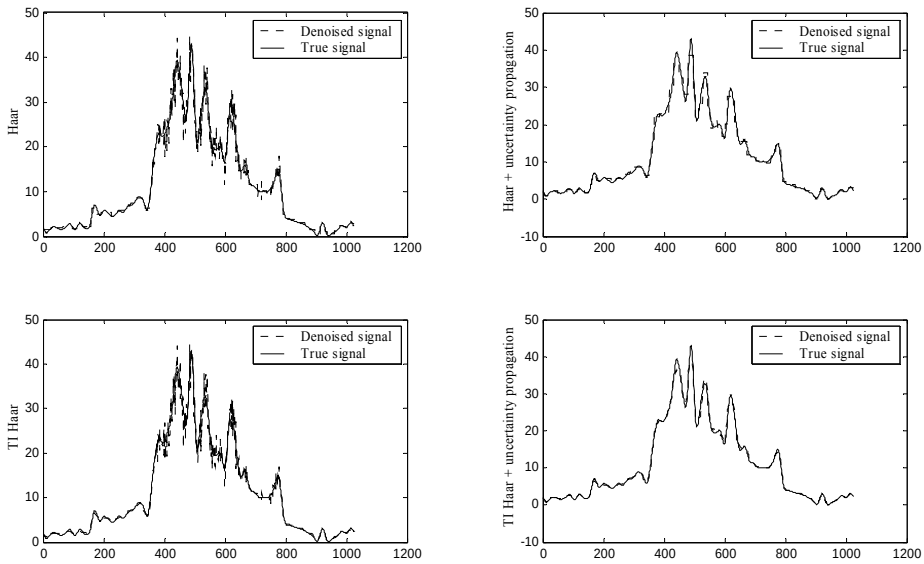


Figure 4.3. Examples of de-noising using the four methods referred in the text ("Haar", "TI Haar", "Haar+uncertainty propagation" and "TI Haar+uncertainty propagation"), for a realization of additive heteroscedastic proportional noise.

4.4 Scale Selection for Data Analysis

When dealing with industrial databases, where hundreds of variables from different points in the plant are being collected, together with product quality variables obtained from the laboratory, it often happens that data sets containing information from all these sources turn out to be quite sparse. This means that they have a lot of "holes", due to variables having different acquisition rates and/or arising from missing data randomly scattered throughout records for each variable, owing to process, instrumentation, communications or data storage related problems. Any efforts directed towards conducting a data analysis task at a very fine time scale (e.g. of minutes), may therefore become useless, for instance, when most of the variables are collected at a coarser time scale (e.g. hours). It would therefore be very appealing, from a practical point of view, to have at our disposal a tool that could suggest what is the finest time scale at which data analysis can be carried out, leaving up to the analyst a final decision about which coarser scale to be in fact adopted.

Method 2 is able to cope both with missing data and data uncertainty, and therefore provides a MRD framework that is instrumental in deciding about a minimum scale for analysis on the basis of either the amount of missing data present or on the uncertainty information available, or even both. After introducing the general methodologies for scale selection in the next subsections (Sections 4.4.1–4.4.3), a real case study is presented, where the proposed approach is applied in on order to select an appropriate scale for conducting data analysis, considering only the presence of missing data (Section 4.4.4), and then another case study is also presented, where the decision is made based upon available uncertainty information (Section 4.4.5).

4.4.1 Scale Selection Based on Missing Data

The four rules underlying implementation of a MRD framework according to *Method 2* (Section 4.1.2) lead to detail and approximation coefficients that can also contain missing data. Let us now define the "reconstruction" procedure that, starting from the coarser approximation coefficients, and using the sets of finer detail coefficients, successively "reconstructs" the finer approximation signals for all the scales below the coarser one, based upon the following pair of rules:

Table 4.2. Rules to be adopted during the reconstruction procedure for the generalized MRD framework (*Method 2*), within the scope of scale selection.

- **Rule 1.** No value missing $\Rightarrow$ use Haar reconstruction procedure;
- **Rule 2.** $\left\{a^{j+1}_{\lceil k/2 \rceil}, d^{j+1}_{\lceil k/2 \rceil}\right\}$ missing $\Rightarrow$ $a^j_k = \text{missing}, d^j_k = \text{missing}$.

By using this "reconstruction" procedure, we come up with a succession of "reconstructed" approximation signals for all scales, which differ from the ones that were obtained during the decomposition phase in the presence of missing data. This happens because when one datum was missing, the decomposition procedure applied rules 2 and 3, introducing a non-missing datum for the coarser approximation coefficient and a zero in the coarser detail coefficient. Then, during the "reconstruction" phase, rule 1 results in two equal non-missing values, where originally we had only one.

Therefore, when there is missing data, the "reconstruction" process "creates" more data (or energy) through a scheme closely related to wavelet interpolation.[24] It is this increase in the energy of the approximation signals at the finest scales, when missing data is present (energy is here defined as the sum of the squares of non-missing values), that allows one to quickly diagnose the scale up to which missing data do play a significant role (interfering with the reconstruction phase), and after which such a behaviour is attenuated.

By plotting the energy of the approximation signals at all scales obtained in the *decomposition procedure*, and that for their counterparts obtained after the *"reconstruction" stage*, along with their difference, to better extract the point where their behaviour stops diverging significantly, a minimum scale to be considered for conducting data analysis can usually be quickly suggested.

[24] This is why the term "reconstruction" is kept between quotation marks, as this scheme does not only reconstruct but also interpolates in the presence of missing data.

4.4.2 Scale Selection Based on Data Uncertainties

Excessive noise may also hinder the analysis at finer scales, because any fine structure that might be present is almost completely immersed under the superimposed unstructured noise component. Basically this means that the true signal's and noise's spectra not only overlap in the higher frequency ranges, but also that the magnitude of the power spectrum for the noise source is sufficiently high in these frequency ranges, so that it disturbs the extraction of accurate frequency information contained in these bands for the true underlying signal.

Therefore, such frequency bands do not convey useful information about the underlying true signal, and should not be used for data analysis. A simple way for identifying uninformative frequency bands consists of applying an uncertainty-based coefficient thresholding methodology, similar to the one presented in Section 4.3, and check whether there is any scale where the detail coefficients are massively thresholded (note that detail coefficients, for a given scale, contain localized information regarding frequency bands, Alsberg et al., 1997). MRD following *Method 3*, that incorporates uncertainty propagation, is adopted for this purpose, and a plot of the energy associated with the original detail coefficients and that for the thresholded ones (or for the difference between them), as well as an additional plot of the percentage of the original energy in that scale that is eliminated by the thresholding operation, will highlight those scales dominated by noise, and therefore not meaningful for performing data analysis.

4.4.3 Scale Selection Based on Missing Data and Data Uncertainties

The suggested procedure for supporting scale selection considering both missing data and data uncertainty results from applying, simultaneously, the two methodologies just presented. Thus, it consists of decomposing data using *Method 2*, after which thresholding is applied to non-missing detail coefficients. The simultaneous analysis of the plots relative to the (differential) distribution of energy contained in the detail coefficients (thresholded according to data uncertainty information) and approximation coefficients (that consider the presence of missing data), as described in Sections 4.4.1–4.4.2, provides the information required to support a decision that considers both missing data as well as data uncertainty.

4.4.4 Case study 1: Scale selection in the Context of Data Analysis Regarding a Pulp and Paper Data Set

A subgroup of nine key quality variables, relative to the pulp produced in an integrated pulp and paper Portuguese mill (Portucel, SA), related to paper structure, strength and optical properties, was collected during four and a half years. These data are to be analysed in order to identify any relevant variation patterns along time, as well as process upsets and disturbances, so that potential root causes can be found and analyzed, leading to process improvement in future operation.

The associated uncertainties were initially estimated using *a priori* knowledge available, regarding measurement devices and the number of significant digits employed in the records (following a Type B procedure for evaluating measurement uncertainty, and assuming constant distributions in ranges defined by the last significant digit; ISO, 1993). However, this approach usually tends to provide rather optimistic estimates for uncertainty figures in industrial settings, since additional noise sources come into the scene when one is not under standard and well controlled conditions. Therefore, these estimates were corrected by also analyzing noise characteristics of the signals using a wavelet-based approach (noise standard deviation was estimated from the details obtained in the first decomposition; Mallat, 1998).

The finest resolution ($j = 0$) present in the data is relative to a daily basis, and the first decision that one has to make concerns a choice of the scale where the analysis should be conducted. This decision was based on a criterion that considers only the presence of missing data, because we knew in advance that this was the major problem with this data set. Therefore, we adopted the methodology described in Section 4.4.1, and analysed all variables in order to decide about the finest scale where such an analysis could be undertaken.

Since the measurement frequencies for all of the nine variables in the plant laboratory are approximately the same, it was not difficult to come up with a single scale that would be valid for all variables. Figure 4.4 illustrates some of the plots thus obtained, relative to variable X_8, where we can clearly see that after scale $j = 3$ (i.e., $2^3 = 8$ days), the effects of the presence of missing data significantly decrease (the energy associated with the approximation coefficients obtained in the decomposition phase gets close to the one obtained after the "reconstruction" phase). Therefore, the scale selected

for conducting data analysis in this case was $j = 3$. This is consistent with the fact that during a period of about one year these variables were measured once a week.

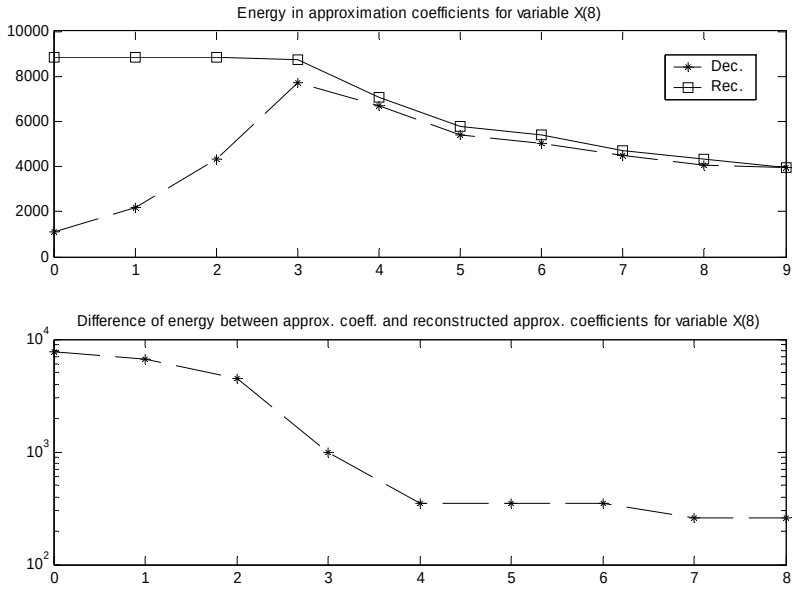


Figure 4.4. (a) Plot of energy contained in the approximation signals after decomposition and reconstruction, at several scales, and (b) semi-log plot of the difference between both of these energies, for each scale.

In order to get insight into the way our MRD framework operates over the data set and the structure of information underlying Figure 4.4, in the next two figures the plots regarding detail coefficients (Figure 4.5) and their associated uncertainties (Figure 4.6) are also presented. As can be clearly seen by the pattern of zero/non-zero detail coefficients, the missing data problem affecting the finer scales is mainly located in a first period of data collection, after which the data acquisition rates were increased.

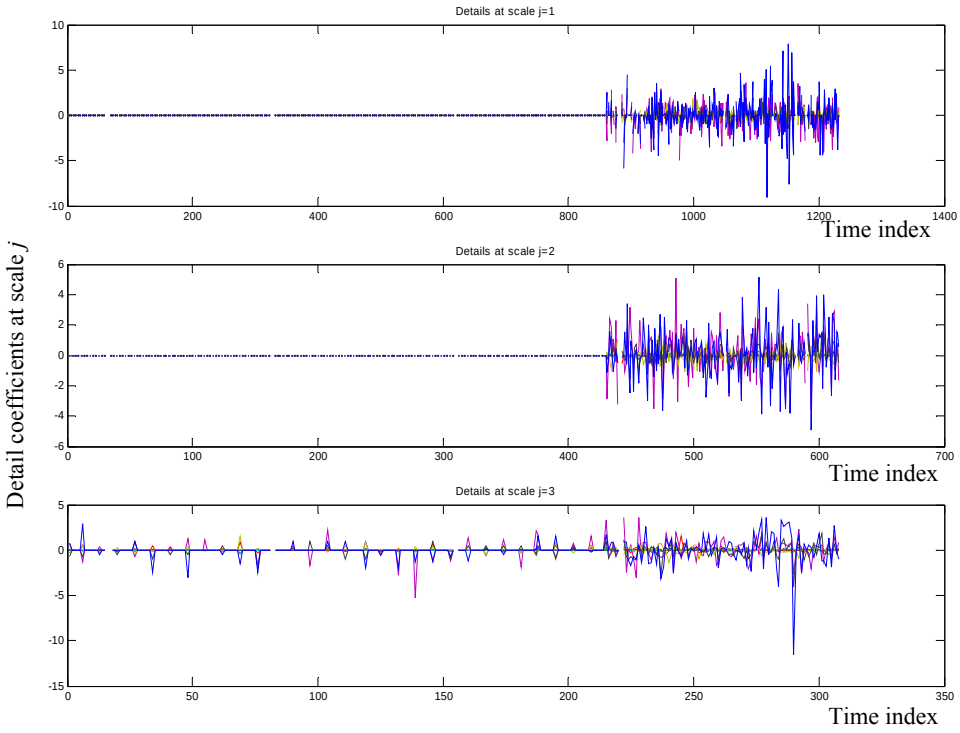


Figure 4.5. Detail coefficients at each scale ($j = 1:3$) obtained by applying our MRD framework (*Method 2*) to the pulp and paper data set.

After selecting the proper scale of analysis ($j = 3$), a modification of the multivariate process monitoring scheme based on PCA is applied. This procedure was designed to take into account the uncertainty information available, along with the approximation coefficients, therefore explicitly considering all available information (raw values and associated uncertainty). Such an approach will be described in Chapter 7, where this example will be completed by performing data analysis at scale $j = 3$.

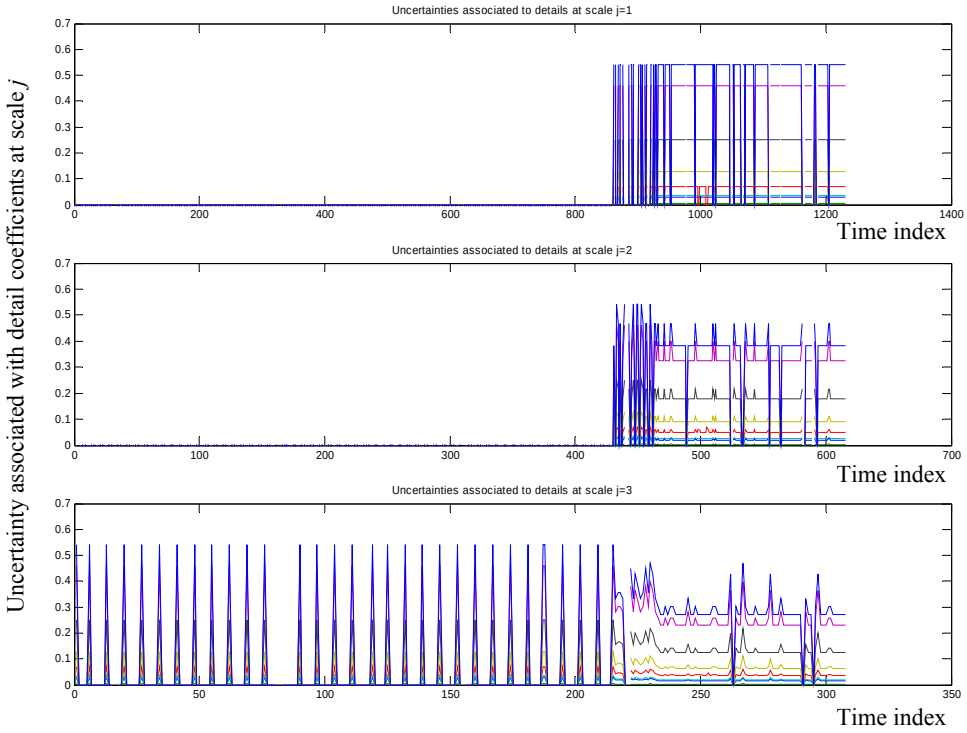


Figure 4.6. Uncertainties associated with the detail coefficients at each scale ($j = 1:3$) obtained by applying our MRD framework (*Method 2*) to the pulp and paper data set.

4.4.5 Case study 2: Analysis of Profilometry Measurements Taken From the Paper Surface

Paper surface plays a key role in its quality, as it is directly connected to a number of important paper properties from the end user's perspective, such as general appearance (optical properties, flatness), printability (e.g. the absorption of ink) and friction. Figure 4.7 presents an accurate surface profile obtained with a mechanical stylus profilometer, where it is quite clear that different surface phenomena are located at different scales: at a coarser scale the presence of waves indicates a problem known as "waviness" or "piping streaks", which have a characteristic wavelength of about $15mm$, while at finer scales, paper micro- and macro-roughness (relative to variations in the cross direction, X, over the ranges of $1\mu m - 100\mu m$ and $100\mu m - 1000\mu m$, respectively) dominate variability. However, there is also an additional contribution to the observed profile that should be considered, due to measurement noise, which is a consequence of the limited

resolution of the measuring device employed for detecting oscillations in the thickness direction (Z) below a certain level, in this case of $8\,nm$.

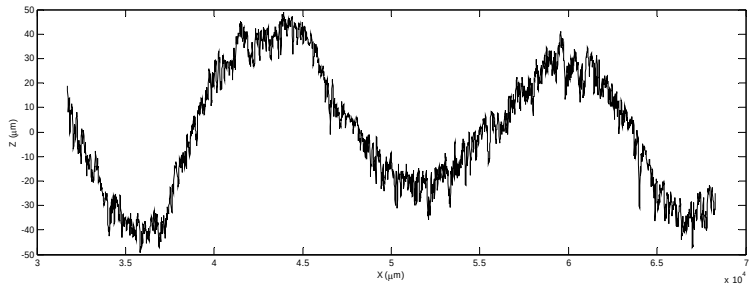


Figure 4.7. Surface profile in the transversal direction, for a paper sample exhibiting waviness phenomena.

The MRD framework based on data uncertainties (*Method 3*) allows for the incorporation of this type of knowledge, and provides important clues about the minimum scale that can be used, as well as scales where the dominant phenomena are located. Figure 4.8 presents the distribution of energy contained in the detail coefficients obtained by decomposing the original profile using a Symmlet-8 wavelet filter (a), and also information regarding the coefficients that are eliminated after applying a thresholding operation that eliminates details below the (propagated) measurement resolution level: the percentage of the energy originally contained in each scale that is removed by the thresholding operation (b) and the percentage of eliminated coefficients at each scale (c).

From Figure 4.8 we can see that the dominating phenomena are located at scale 11, corresponding to $8.93\times 2^{11}\,\mu\text{m}\cong 18.3\,\text{mm}$ (the separation between two successive samples in the X direction is of about $8.93\ \mu\text{m}$), i.e., quite close to the characteristic wavelength of the waviness phenomena, and that the profile is almost unaffected by measurement noise at all scales, as only very few coefficients are discarded at the finest scales as a consequence of the limited resolution of the measuring device.

Therefore, the high resolution profilometer is indeed suitable to assess the fine details of paper surface (minimum scale for analysis is $j=1$), and one may also conclude that, in this particular case, all the scales do contain potentially relevant information regarding

the characterization of underlying phenomena. Furthermore, efforts for analysing and monitoring waviness should focus mainly around scales $j = 10$ and $j = 11$.

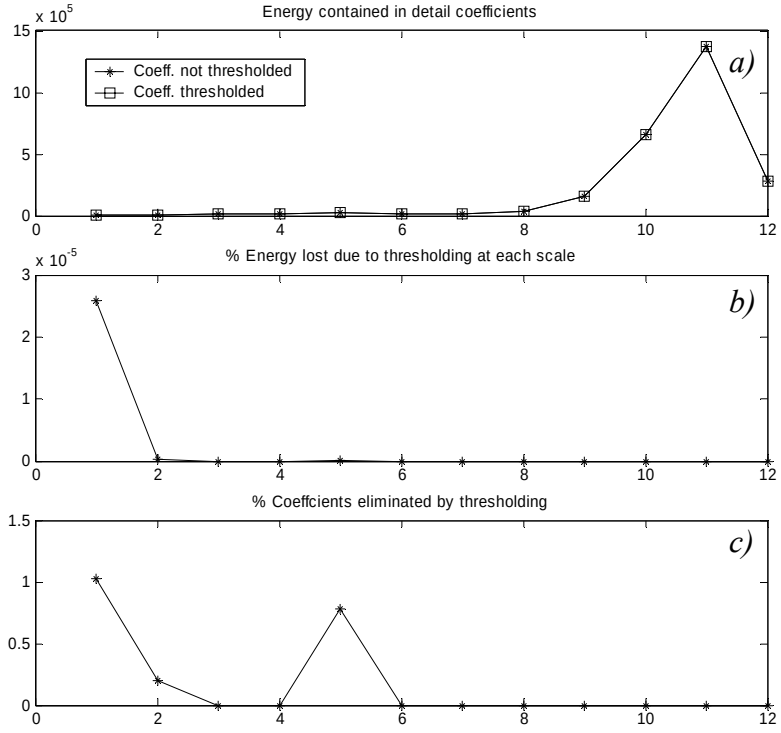


Figure 4.8. Plots of (a) distribution of energy in detail coefficients across scales, (b) percentage of energy originally contained in each scale that is removed by the thresholding operation (relatively to the original energy content of that scale) and (c) percentage of eliminated coefficients in each scale (relatively to the original number of coefficients in that scale).

4.5 Conclusions

MRD frameworks play an essential role when one needs to focus data analysis at a particular scale or to identify the several contributions to overall phenomenon, arising from different scales in time or length. However, the presence of missing data raises serious difficulties in implementing classical MRD based on wavelets. The incorporation of data uncertainty in the analysis is also desirable from the standpoint of using all the available information right from the beginning. Therefore, in this chapter, three MRD frameworks were proposed that provide a basis for handling such issues, and guidelines about their use were also put forward.

Methods 1 and *2* handle the presence of missing data and any structure of data uncertainties, the former being especially devoted to piecewise constant signals. *Method 3* handles those cases where no missing data is present, incorporating data uncertainty in the computation of detail and approximation coefficients.

It should be stressed that *Methods 1* and *2* are not extensions of the wavelet transform in a strict sense, as some of their fundamental properties do not always hold, such as the energy conservation property in *Method 2* (in the sense of Parseval formula, Mallat, 1998). However, they allow one to extend the wavelet multiresolution decomposition to contexts where it could not be applied otherwise (at least without some serious data pre-processing efforts), namely when we have missing data. Furthermore, such methods provide new tools for addressing other types of problems in data analysis, such as that of selecting a proper scale for analysis. Several simulated and real world problems illustrate the use of the various methodologies suggested here and their practical potential value.

Chapter 5. Integrating Data Uncertainty Information in Regression Methodologies

The MRD frameworks presented in the previous chapter generate new data sets, with detail and approximation coefficients, together with their associated uncertainties. On the other hand, the availability of uncertainty information for data arising from many different sources is increasing rapidly, as a consequence of the efforts undertaken in the fields of metrology and standardization, namely regarding the characterization and quantification of measurement uncertainty, in a rigorous and normalized way (ISO, 1993). This implies that there is not only one data table to be explored, but rather two tables: the usual raw data table, and another one with the associated uncertainties.

Therefore, in order to take full advantage of all the available information, data analysis tools should also explicitly consider data uncertainty information in their formulations, in order to become more flexible, in the sense of being adequate for application in situations encompassing a wider diversity of measurement error structures, including those whose measurement error structures are not covered by more conventional techniques.

In fact, the majority of conventional techniques commonly applied to chemical processes, do rely on simplified assumptions regarding the nature of errors included in their general statistical model structures, not taking explicitly and quantitatively into account data quality, or only doing so in an implicit or tacit way. More specifically, the error term is normally considered as arising from several different sources, such as

modelling mismatch (inadequate model structure assumed), *uncontrolled interferences* (Martens & Naes, 1989) and *measurement noise*, and their statistical descriptions are based on an assumed homoscedastic behaviour (i.e., with constant variance). This may be a reasonable assumption for the two first error sources (modelling mismatch and uncontrolled interferences), for which we are not usually able to provide additional a priori knowledge regarding their behaviour along time, but that is not necessarily so for measurement noise, for which, furthermore, the increasing availability of measurement uncertainty information can be explored.

Some examples of application contexts where uncertainty-based methods can become quite useful, include the analysis of spectra (that often present noise, not rarely of an heteroscedastic nature, and in the presence of strong correlations in the predictors), microarray data (where heteroscedasticity is mainly due to different levels of colour definition in the spotted arrays), laboratorial data (where measurements of quality variables are often correlated and affected by different levels of uncertainties) and industrial data.

Therefore, it is quite appropriate and timely to develop and apply methods that take into account explicitly and consistently this important piece of information, and in this chapter we address this issue within the scope of regression methodologies, given their importance and generalized use in the analysis of industrial data. Then, in the next chapter, we also refer how this type of information can be used in process optimization, to come up with better operation policies, and in Chapter 7 we address its integration in multivariate statistical process control.

In the next section, we refer several linear regression methodologies, ranging from conventional techniques to those designed to take into account data uncertainties (multivariate least squares, maximum likelihood principal components regression), and others whose potential to deal with noisy data is well known (partial least squares, principal components regression and ridge regression), as well as modifications of these methods that were developed in the context of this thesis. Then, in the following section we present two case studies, that provide the ground for comparing all the methods considered. In the third section, main results and some computational issues are discussed, with final conclusions drawn in the fourth section.

5.1 Multivariate Linear Regression Methods

This section is devoted to the description of four groups of multivariate linear regression methods that have the potential to accommodate measurement noise information, either explicitly or implicitly. As already referred, our focus on multivariate linear regression arises from the quite widespread use for this type of approaches in the development of input/output models for industrial and/or laboratorial applications. The several methodologies here addressed are clustered under four separate groups, according to their affinity: ordinary least squares (OLS), ridge regression (RR), principal components regression (PCR) and partial least squares (PLS). Besides these four basic methods, that do not explicitly incorporate measurement uncertainty information, several alternatives already developed are also presented, as well as other modifications proposed here, that do take uncertainty information explicitly into consideration.

5.1.1 OLS Group

The Ordinary Least Squares (OLS) (Draper & Smith, 1998) and Multivariate Least Squares (MLS) (Martínez *et al.*, 2002a; Río *et al.*, 2001) parameter estimates for a linear regression model are given by the solutions of the optimization problems formulated in equations (5.1)-(5.2) of Table 5.1.

Table 5.1. Formulation of optimization problems underlying OLS, MLS and MLMLS methods.[25]

OLS	$\hat{b}_{OLS} = \underset{b=\left[b_0 \cdots b_p\right]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y(i) - \hat{y}(i) \right)^2 \right\}$	(5.1)
MLS	$\hat{b}_{MLS} = \underset{b=\left[b_0 \cdots b_p\right]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \frac{\left(y(i) - \hat{y}(i) \right)^2}{s_e^2(i)} \right\}$	(5.2)
MLMLS	$\hat{b}_{MLMLS} = \underset{b=\left[b_0 \cdots b_p\right]^T}{\arg\max} \Lambda(b)$ $\Lambda(b) = -\frac{1}{2} n \ln\left(2\pi\right) - \sum_{i=1}^{n} \ln\left(\sigma_{\varepsilon_i}\right) - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{\left(y(i) - \hat{y}(i) \right)^2}{\sigma_{\varepsilon_i}^2} \right)$	(5.3)

[25] $\hat{y}(i) = b_0 + X(i,:)b$, where $X(i,:)$ is the i[th] row (observation) of the matrix containing all predictor variables in its columns, X.

OLS tacitly assumes a homoscedastic behaviour (i.e., with constant variance) for the noise error term in the standard linear regression model:

$$y(i) = b_0 + b_1 X_1(i) + .. + b_p X_p(i) + \varepsilon(i) \tag{5.4}$$

On the other hand, MLS is built upon an Error in Variables (EIV) functional relationship, relating true values of both the input and output variables, which are then affected by zero mean random errors ($\Delta\eta(i)$ and $\Delta\xi_j(i)$), with a given covariance structure (assumedly known):

$$\eta(i) = b_0 + b_1 \xi_1(i) + .. + b_p \xi_p(i) \tag{5.5}$$

$$\begin{aligned} y(i) &= \eta(i) + \Delta\eta(i) \\ X_j(i) &= \xi_j(i) + \Delta\xi_j(i) \end{aligned} \tag{5.6}$$

In the denominator of equation (5.2) we can find a term, $s_e^2(i)$, that results from the summation of the uncertainties associated with the response to the ones arising from the propagation of uncertainties of the predictors to the response, according to a formula derived from error propagation theory (Lira, 2002; Martínez *et al.*, 2002a):

$$s_e^2(i) = uy(i)^2 + \sum_{j=1}^{p} \hat{b}_j{}^2 uX(i,j)^2 + 2\sum_{j=2}^{p}\sum_{k=j+1}^{p} \hat{b}_j \hat{b}_k \operatorname{cov}\left(\Delta\xi_j(i), \Delta\xi_k(i)\right) \tag{5.7}$$

where $uX(i,j)$ and $uy(i)$ are the uncertainties associated with the i^{th} observation of the j^{th} input and output variables, respectively, and $\Delta\xi_j(i)$ is the random error affecting the i^{th} measurement of variable j; $\hat{b}_j$ represents the coefficient of the linear regression model associated with variable j. By imposing the necessary optimality conditions for local optima to (5.2), it is possible to set up an algorithmic procedure for the numerical

solution of the optimization problem underlying MLS, such that, at each iteration, a new estimate for parameter vector b is provided, through the solution of a system of $m+1$ linear equations of the type $Rb=g$ (Lisý *et al.*, 1990; Martínez *et al.*, 2002a):

$$\overbrace{\begin{bmatrix} \sum_{i=1}^{n}\frac{1}{s_e^2(i)} & \sum_{i=1}^{n}\frac{X(i,2)}{s_e^2(i)} & \cdots & \sum_{i=1}^{n}\frac{X(i,m+1)}{s_e^2(i)} \\ \sum_{i=1}^{n}\frac{X(i,2)}{s_e^2(i)} & \sum_{i=1}^{n}\frac{X(i,2)^2}{s_e^2(i)} & \cdots & \sum_{i=1}^{n}\frac{X(i,m+1)\times X(i,2)}{s_e^2(i)} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n}\frac{X(i,m+1)}{s_e^2(i)} & \sum_{i=1}^{n}\frac{X(i,m+1)\times X(i,2)}{s_e^2(i)} & \cdots & \sum_{i=1}^{n}\frac{X(i,m+1)^2}{s_e^2(i)} \end{bmatrix}}^{R} \overbrace{\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}}^{b} = \underbrace{\begin{bmatrix} \sum_{i=1}^{n}\frac{y_i}{s_e^2(i)} \\ \sum_{i=1}^{n}\left[\frac{y_i\times X(i,2)}{s_e^2(i)}+\frac{1}{2}\left(\frac{e(i)}{s_e^2(i)}\right)^2 2b_1 uX(i,2)^2\right] \\ \vdots \\ \sum_{i=1}^{n}\left[\frac{y_i\times X(i,m+1)}{s_e^2(i)}+\frac{1}{2}\left(\frac{e(i)}{s_e^2(i)}\right)^2 2b_{m+1} uX(i,m+1)^2\right] \end{bmatrix}}_{g} \tag{5.8}$$

where $e(i)=y(i)-\hat{y}(i)$.

The method whose objective function is presented in Table 5.1 through equation (5.3) is derived from the analysis of the Berkson case ("controlled regressors with error"), within the scope of EIV models (Mandel, 1964; Seber & Wild, 1989), and under the assumption of Gaussian errors. The objective function arises from the maximization of the likelihood function thus obtained, and this approach was included in our present study given the similarity between the quadratic functional part of its objective function and the one underlying MLS, as well as due to its simplicity. As the solution for the Berkson case formulation is in some sense similar to MLS (Seber & Wild, 1989), we call to the above formulation Maximum Likelihood Multivariate Least Squares (MLMLS), to stress the statistical motivation of the underlying objective function. More information regarding this method can be found in Appendix A.

5.1.2 RR Group

A well known characteristic of the OLS method is the fact that the variance of its parameter estimates increases when the input variables get correlated. Computational simulations showed us that the same applies to MLS. One possible way to address this issue consists of enforcing an effective shrinkage in the coefficients under estimation, following a ridge regression (RR) regularization approach. It basically consists of adding an extra term to the objective function that penalizes large solutions (in a square norm sense). Optimization formulations underlying RR estimates (Draper & Smith, 1998; Hastie *et al.*, 2001), as well as those proposed for its counterparts based on MLS and MLMLS, rMLS and rMLMLS, respectively (standing for "ridge MLS" and "ridge MLMLS"), are presented in Table 5.2.

Table 5.2. Formulation of optimization problems underlying RR, rMLS and rMLMLS.

RR	$\hat{b}_{RR} = \underset{b=\left[b_0 \cdots b_p\right]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y(i) - \hat{y}(i) \right)^2 + \lambda \sum_{j=1}^{p} b(j)^2 \right\}$	(5.9)
rMLS	$\hat{b}_{rMLS} = \underset{b=\left[b_0 \cdots b_p\right]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \frac{\left(y(i) - \hat{y}(i) \right)^2}{s_e^2(i)} + \lambda \sum_{j=1}^{p} b(j)^2 \right\}$	(5.10)
rMLMLS	$\hat{b}_{rMLMLS} = \underset{b=\left[b_0 \cdots b_p\right]^T}{\arg\min} \left\{ \sum_{i=1}^{n} \ln\left(s_e(i) \right) + \sum_{i=1}^{n} \frac{\left(y(i) - \hat{y}(i) \right)^2}{s_e^2(i)} + \lambda \sum_{j=1}^{p} b(j)^2 \right\}$	(5.11)

It can be shown that, just as the shrinkage term, $\lambda \sum_{j=1}^{p} b(j)^2$, stabilizes the inversion step of OLS in RR (improving the condition of matrix $X^T X$ by adding a positive constant, λ, to the diagonal elements), it also stabilizes MLS' R matrix in a similar way (except for the first row, where no constant λ is added in the first entry).

5.1.3 PCR Group

PCR (Jackson, 1991; Martens & Naes, 1989) is another methodology for handling collinearity among predictor variables. It uses those uncorrelated linear combinations of the input variables that explain most of the input space variability (from PCA, Appendix D) as the new set of predictors, where the response is to be regressed onto. These predictors are orthogonal, and therefore the collinearity problem is overcome if we

disregard the linear combinations with small variability explanation power (Martens & Mevik, 2001).

After developing MLPCA, which estimates the PCA subspace in an optimal maximum likelihood sense, when data are affected by measurement errors with a known uncertainty structure (Wentzell *et al.*, 1997a), Wentzell *et al.* (1997b) applied it in the context of developing a PCR methodology that incorporates measurement uncertainties (MLPCR). As in PCR, MLPCR consists of first estimating a PCA model, now using MLPCA, in order to calculate the scores through non-orthogonal (maximum likelihood) projections to the estimated MLPCA subspace (instead of the PCA's orthogonal projections), and then applying OLS to develop a final predictive model. This technique makes use of the available uncertainty information in the former phases (estimation of a MLPCA model and calculation of its scores), but not during the stage at which OLS is applied. Therefore, Martínez *et al.* (2002) proposed a modification to the regression phase, in order to make it consistent with the efforts of integrating uncertainty information carried out in the initial stages, that consists of replacing OLS by MLS (we will call this modification as MLPCR1). In order to implement MLS in the second phase, estimated score uncertainties for the i^{th} observation need to be calculated, given by the diagonal elements of the following matrix (Martínez *et al.*, 2002a):

$$Z_i = \left\{ P^T \left[diag\left(uX(i,:) \right) \right]^{-1} P \right\}^{-1} \tag{5.12}$$

where *diag* is an operator that converts a vector into a diagonal matrix, and P is the matrix of maximum likelihood loadings.

In the present work, these algorithms based on OLS and MLS (MLPCR and MLPCR1, respectively), are also compared with the one obtained using the MLMLS algorithm, instead of OLS, during the second phase of MLPCR (MLPCR2).

5.1.4 PLS Group

PLS (Geladi & Kowalski, 1986; Haaland & Thomas, 1988; Helland, 1988, 2001b; Höskuldsson, 1996; Jackson, 1991; Martens & Naes, 1989; Wold *et al.*, 2001) is a widely used algorithm in the chemometrics community, that also adequately handles

noisy data with correlated predictors in the estimation of a linear multivariate model. As in PCR, PLS finds a set of uncorrelated linear combinations of the predictors, belonging to some lower dimensional subspace in the X-variables space, where *y* is to be regressed onto. However, in PLS this subspace is the one that, while still covering well the X-variability, provides a good description of the variability exhibited by the Y-variable(s). Here we will make reference to a pair of classes of PLS algorithms, one implemented from *raw data*, and another based upon *covariance matrices*.

PLS algorithms implemented directly from raw data

The algorithmic nature of PLS (Geladi & Kowalski, 1986; Höskuldsson, 1996) can be translated into the solutions of a succession of optimization sub-problems (Haaland & Thomas, 1988; Jackson, 1991; Martens & Naes, 1989), as presented in the first column of Table 5.3 for one of its common versions, relative to the case of a single response variable. However, if besides having available raw data, $[X \mid y]$, we also know their respective uncertainties, $[uX \mid uy]$, then one way to incorporate this additional information into a PLS algorithm is through an adequate reformulation of the optimization sub-tasks. Therefore, we have modified the objective functions underlying each optimization sub-problem in order to incorporate measurement uncertainties, but still preserving the successful algorithmic structure of PLS. Such a sequence of optimization sub-problems is presented in the second and third columns of Table 5.3, where MLS and MLMLS replace OLS in the algorithmic stages, giving rise to the uncertainty-based counterparts, uPLS1 and uPLS2, respectively. More details regarding these methods are presented below.

i) Computation of the X-scores vector (t)

Computation of the X-scores vector, for each dimension, involves solving the optimization problem formulated in step 3 of Table 5.3. Its analytic solution can be derived using multivariate calculus (Magnus & Neudecker, 1988), leading to equation (5.13), but it provides the same numerical results as the maximum likelihood projection formula for computing the X-scores in MLPCA presented in (Wentzell *et al.*, 1997b):

$$t = vec(X) \cdot \Omega \cdot \left(w \otimes I_n \right) \cdot \left[\left(w \otimes I_n \right)^T \cdot \Omega \cdot \left(w \otimes I_n \right) \right]^{-1} \tag{5.13}$$

where Ω is a diagonal matrix with the inverses of the elements of the *vectorized* matrix uX along the diagonal, I_n is the identity matrix of dimension $n \times n$, $\otimes$ is the Kronecker product operator and *vec* is the operator that *vectorizes* higher order tensors [21].

Another issue in the calculation of the X-scores is related with the computation of the associated uncertainties. In uPLS1 and uPLS2 uncertainties were propagated to the scores under the assumption of negligible uncertainties in the weights or loadings (a more complete treatment can be built around the results of Goodman and Haberman, 1990). As the scores can be given as maximum likelihood projections onto the subspace spanned by the weight vector, we can use an expression similar to (5.12) in order to calculate uncertainty propagated to the a^{th} X-scores. Furthermore, measurement errors affecting variables are also assumed to be statistically independent.

ii) Computation of X-weights (w) and X-loadings (p) vectors

In the computation of the X-weights vector, our optimization problem can be seen as a succession of univariate regression problems of the y-score, u, onto $X(:,j)$ (the j^{th} column of X), with zero intercept. However, as both u and $X(:,j)$ have associated uncertainties, the adequate way to estimate the $w(j)$ coefficient, in the sense of the optimization sub-task formulated in step 2, is by means of BLS or MLMLS (without intercept). The same applies to the calculation of the X-loadings, where BLS/MLMLS are now applied to the regression of t onto $X(:,j)$, with the score uncertainties calculated as referred above and the X uncertainties provided as inputs or calculated for the residual matrices, obtained after deflation, as shown below.

iii) Computation of uncertainties for the X and y residual matrices

After deflation, in order to carry on with uPLS1 and uPLS2, we need to update the uncertainties associated with residual matrices E_a and F_a, which play, for a>1, the same role that X and Y have played during the calculations for a=1. This can be done by applying error propagation theory (once again, we have assumed that only the scores do carry significant uncertainties).

PLS algorithms implemented from covariance matrices

There are several alternative ways for developing a PLS model, most of them leading to very similar or even exactly the same results. In fact, Helland (1988) has shown the

equivalence between two of such algorithms (one based on orthogonal scores and another using orthogonal loadings instead), both of them based on available raw data matrices for the predictors and response variables. Another class of PLS methods, that encompasses the so called SIMPLS, developed by Sijmen de Jong (see Table 5.4), or the approach presented by Kaspar & Ray (1993) built upon previous work from Höskuldsson (1988), consists of algorithms entirely based on data covariance or cross-product matrices. For the single response case, a SIMPLS solution provides exactly the same results as Svant Wold's orthogonalized PLS algorithm, leading to only minor differences when several outputs are considered. Matrices *S* and *s* in Table 5.4 do play a central role in PLS. A theoretical analysis of this algorithm (Helland, 1988; Phatak, 1993) leads to the conclusion that the calculated vector of coefficients, when *a* latent variables are considered, $\hat{\beta}_{PLS}^{a}$, is given by:

$$\hat{\beta}_{PLS}^{a} = V_a \left(V_a{}^T S V_a \right)^{-1} V_a{}^T s \tag{5.14}$$

where $V_a = \left[v_1, v_2, \cdots, v_a \right]$ is any ($m \times a$) matrix whose columns span the Krylov subspace $\Re^a \left(s; S \right)$, i.e., the subspace generated by the first *a* columns of the Krylov sequence, $\left\{ s, Ss, \cdots, S^{a-1} s \right\}$. Thus, matrices *S* and *s* define the structure of the relevant Krylov subspace where a PLS solution lies. In fact, the columns of the PLS weighting matrix, *W*, that define the subspace of the full predictor space with maximal covariance with the response, do form an orthogonal base of $\Re^a \left(s; S \right)$.

Table 5.3. PLS as a succession of optimization sub-problems (first column), and its counterparts, that make use of information regarding measurement uncertainty.

PLS	uPLS1	uPLS2
Step 1. Pre-treatment Center X and y; Scale X and y.	***Step 1. Pre-treatment*** Center X and y ; Scale X and y. Scale X and y uncertainties.	***Step 1. Pre-treatment*** Center X and y; Scale X and y. Scale X and y uncertainties.
Begin For Cycle a=1 : # latent variables	**Begin For Cycle** a=1 : # latent variables	**Begin For Cycle** a=1 : # latent variables
Step 2. Calculate the a^{th} X-weights vector (w) $w = \arg\min_{w} \sum_{i=1}^{n}\sum_{j=1}^{m}\left(X(i,j) - u(i)\times w(j)\right)^2$ $w_{new} \leftarrow w_{old} / \left\|w_{old}\right\|$ Note: for a=1, the Y-scores, u, are equal to y.	***Step 2. Calculate the a^{th} X-weights vector (w)*** $w = \arg\min_{w} \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(X(i,j) - u(i)\times w(j)\right)^2}{uX(i,j)^2 + w(j)^2 \times uy(i)^2}$ $w_{new} \leftarrow w_{old} / \left\|w_{old}\right\|$	***Step 2. Calculate the a^{th} X-weights vector (w)*** $w(j) = \arg\min_{w(j)} \left\{ -\frac{1}{2} n \ln(2\pi) - \sum_{i=1}^{n} \ln\left(\sigma_{\varepsilon_{i,j}}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\left(X(i,j) - \hat{X}(i,j)\right)^2}{\sigma^2_{\varepsilon_{i,j}}} \right) \right\}$, where, $\sigma^2_{\varepsilon_{i,j}} = \left(uX(i,j)\right)^2 + \left(uu(i)\right)^2 w(j)^2$; $w_{new} \leftarrow w_{old} / \left\|w_{old}\right\|$
Step 3. Calculate a^{th} X-scores vector (t) $t = \arg\min_{t} \sum_{i=1}^{n}\sum_{j=1}^{m}\left(X(i,j) - t(i)\times w(j)\right)^2$	***Step 3. Calculate a^{th} X-scores vector (t)*** $t = \arg\min_{t} \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(X(i,j) - t(i)\times w(j)\right)^2}{uX(i,j)^2}$	***Step 3. Calculate a^{th} X-scores vector (t)*** $t = \arg\min_{t} \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(X(i,j) - t(i)\times w(j)\right)^2}{uX(i,j)^2}$
Step 4. Calculate a^{th} X-loadings vector (p) $p = \arg\min_{p} \sum_{i=1}^{n}\sum_{j=1}^{m}\left(X(i,j) - t(i)\times p(j)\right)^2$	***Step 4. Calculate a^{th} X-loadings vector (p)*** $p = \arg\min_{p} \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(X(i,j) - t(i)\times p(j)\right)^2}{uX(i,j)^2 + p(j)^2 \times ut(i)^2}$	***Step 4. Calculate a^{th} X-loadings vector (p)*** $p(j) = \arg\min_{p(j)} \left\{ -\frac{1}{2} n \ln(2\pi) - \sum_{i=1}^{n} \ln\left(\sigma_{\varepsilon_{i,j}}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\left(X(i,j) - \hat{X}(i,j)\right)^2}{\sigma^2_{\varepsilon_{i,j}}} \right) \right\}$, where $\sigma^2_{\varepsilon_{i,j}} = \left(uX(i,j)\right)^2 + \left(ut(i)\right)^2 p(j)^2$
Step 5. Re-scale X-loadings, X-scores and X-weights $p_{new} \leftarrow p_{old} / \left\|p_{old}\right\|$; $t_{new} \leftarrow t_{old} \times \left\|p_{old}\right\|$; $w_{new} \leftarrow w_{old} \times \left\|p_{old}\right\|$	***Step 5. Re-scale X-loadings, X-scores and X-weights*** $p_{new} \leftarrow p_{old} / \left\|p_{old}\right\|$; $t_{new} \leftarrow t_{old} \times \left\|p_{old}\right\|$; $w_{new} \leftarrow w_{old} \times \left\|p_{old}\right\|$ *Step 5.1. Update ut(i), i=1:n.*	***Step 5. Re-scale X-loadings, X-scores and X-weights*** $p_{new} \leftarrow p_{old} / \left\|p_{old}\right\|$; $t_{new} \leftarrow t_{old} \times \left\|p_{old}\right\|$; $w_{new} \leftarrow w_{old} \times \left\|p_{old}\right\|$ *Step 5.1. Update ut(i), i=1:n.*
Step 6. Regression of u on t (b) $b = \arg\min_{b} \sum_{i=1}^{n}\left(u(i) - t(i)\times b\right)^2$.	***Step 6. Regression of u on t (b)*** $b = \arg\min_{b} \sum_{i=1}^{n}\frac{\left(u(i) - b\times t(i)\right)^2}{uu(i)^2 + b^2 \times ut(i)^2}$	***Step 6. Regression of u on t (b)*** $b = \arg\min_{b} \left\{ -\frac{1}{2} n \ln(2\pi) - \sum_{i=1}^{n} \ln\left(\sigma_{\varepsilon_{i,j}}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{\left(u(i) - \hat{u}(i)\right)^2}{\sigma^2_{\varepsilon_{i,j}}} \right) \right\}$, where $\sigma^2_{\varepsilon_{i,j}} = \left(uu(i)\right)^2 + \left(ut(i)\right)^2 b^2$
Step 7.Calculation of X and Y residuals $E_a = E_{a-1} - t_a p_a^T \ (X = E_0)$ $F_a = F_{a-1} - b_a t_a \ (y = F_0)$ Note: Continue the calculations with E_a playing the role of X and F_a the one of y(u).	***Step 7.Calculation of X and Y residuals*** $E_a = E_{a-1} - t_a p_a^T \ (X = E_0)$ $F_a = F_{a-1} - b_a t_a \ (y = F_0)$ *Step 7.1. Up-date* $\left\{uE(i,j),\ uF(i)\right\}_{i=1,n;j=1,m}$.	***Step 7.Calculation of X and Y residuals*** $E_a = E_{a-1} - t_a p_a^T \ (X = E_0)$ $F_a = F_{a-1} - b_a t_a \ (y = F_0)$ *Step 7.1. Up-date* $\left\{uE(i,j),\ uF(i)\right\}_{i=1,n;j=1,m}$.
End For Cycle	**End For Cycle**	**End For Cycle**

Table 5.4. SIMPLS algorithm (de Jong *et al.*, 2001).

$S = X^T X$
$s = X^T y$
for a=1,...,A
  r=1st left singular vector of s
  $r = r/(r^T S r)^{1/2}$
  $R = [R, r]$
  $P = [P, Sr]$
  $s = \left[I - P\left(P^T P\right)^{-1} P^T \right] s$
end
$T = XR$

The relevancy of S and s for PLS provided the necessary motivation to direct some efforts towards the incorporation of uncertainty information in the computation of better *estimates* for both of these matrices. The reason why we have not called them estimates until now is due to the lack of a consistent statistical population model underlying PLS (Helland, 2001a, 2001b, 2002). However, when we now say that our goal is to calculate "better" covariance matrices, this implies that some goodness criteria must be assumed. Therefore, in order to give a step forward, towards the integration of measurement uncertainties in our analysis, one should postulate a statistical model, in order to provide an estimation setting for the covariance matrices S and s. For the sake of the present work, we consider the following latent variable multivariate linear relationship for $Z = \left[x^T \mid y \right]^T$, that has the ability of incorporating heteroscedastic measurement errors with known uncertainties (these uncertainties are considered by now to be independent of the true levels for the noiseless measurands):

$$Z(k) = \mu_Z + A \cdot l(k) + \varepsilon_m(k) \tag{5.15}$$

where Z is the $(m+1)\times 1$ vector of measurements, μ_Z is the $(m+1)\times 1$ mean vector of x, A is the $(m+1)\times a$ matrix of model coefficients, l is the $a\times 1$ vector of latent variables and ε_m is the $(m+1)\times 1$ vector of measurement noise. The probability density functions assumed for each random component are:

$$\begin{aligned} &l(k) \sim iid\ MN_a\left(0, \Delta_l\right) \\ &\varepsilon_m(k) \sim id\ MN_{m+1}\left(0, \Delta_m(k)\right) \\ &l(k) \text{ and } \varepsilon_m(j) \text{ are independent } \forall k, j \end{aligned} \tag{5.16}$$

where *MN* stands for multivariate normal distribution, Δ_l is the covariance matrix of the latent variables, $\Delta_m(k)$ is the covariance matrix of the measurement noise at time *k*, given by $\Delta_m(k) = diag\left(\sigma_m^2(k)\right)$.

Thus, for estimating the covariance matrix, we assume a multivariate behaviour for *Z* that can be adequately described by propagation of the underlying variation of *p* latent variables, plus added noise in the full variable space. This model, and the calculation details associated with the estimation of the unknown parameters, will be described in more detail in Chapter 7.

Under the conditions stated above, the probability density function of *Z* is a multivariate normal distribution with the following form:

$$Z(k) \sim id\ MN_{m+1}\left(\mu_Z, \Sigma_Z(k)\right) \tag{5.17}$$

where

$$\begin{aligned} &\Sigma_Z(k) = \Sigma_l + \Delta_m(k) \\ &\Sigma_l = A\,\Delta_l\,A^T \end{aligned} \tag{5.18}$$

With the raw measurements (*Z*) and the associated uncertainties (from which we can calculate $\Delta_m(k)$), it is possible to estimate μ_Z and Σ_l through the maximization of the associated likelihood function (Chapter 7). Matrix $\Sigma_Z(k) = \Sigma_l + \Delta_m(k)$ is the estimate of the covariance matrix for noisy measurements at time step *k*, but as PLS is based on *S* and *s*, it requires single estimates for the population parameters (and not one per time step *k*). Thus, we keep the estimate of the covariance for noiseless data, $\hat{\Sigma}_l$, but average

out the heteroscedastic square uncertainties, in order to come up with a single term, $\overline{\hat{\Delta}}_m$, leading to:

$$\hat{\Sigma}_Z \cong \hat{\Sigma}_l + \overline{\hat{\Delta}}_m \tag{5.19}$$

With the estimate of Σ_Z, we can finally calculate the estimates for S and s: $S = \hat{\Sigma}_Z(1:m,1:m)$, $s = \hat{\Sigma}_Z(1:m,m+1)$. The algorithm that consists of implementing the SIMPLS algorithm, with these matrices as inputs, will be here referred to as uPLS3.

In the present context, the full measurement space is used in order to estimate Σ_Z ($a = m$), so that the PLS algorithm can be used to compute the relevant subspace for prediction, instead of doing so at an earlier estimation stage. In the *prediction phase*, when new values for the predictors become available along with their measurement uncertainties, and the goal is to predict what the value of the response variable would be, we add an additional calculation step *before* applying the uPLS3 regression vector (calculated in the *estimation phase*). This step consists of projecting the new multivariate observation in the full *X*-space onto the subspace that is relevant for predictions (i.e., the one spanned by the columns of the weighting matrix, *W* in PLS *or* *R* in SIMPLS). The availability of the associated uncertainties leads to a generally non-orthogonal projection methodology that consists of estimating the projected points using a maximum likelihood approach, just as the one adopted in MLPCA (Wentzell *et al.*, 1997b).

In the present work, another algorithm was also developed and tested, that implements the same non-orthogonal projection operation, but using the weighting matrix provided by PLS (an hybrid version of classic PLS, since it contains a projection step that incorporates measurement uncertainty), here referred to as uPLS4. For the sake of completeness, we also introduced another methodology, based on the same weighting matrix as uPLS3, but that bypasses the non-orthogonal projection step, designated as uPLS5.

5.2 Monte Carlo Simulation Comparative Study

In this section we present the results reached from a comparative analysis encompassing all the methods mentioned above (PLS, uPLS1, uPLS2, uPLS3, uPLS4, uPLS5, RR, rMLS, rMLMLS, PCR, MLPCR, MLPCR1, MLPCR2, OLS, MLS and MLMLS).

Case studies 1 and 2 provide different contexts to set the ground for comparing multivariate linear regression methods. In both of them, a latent variable model structure is adopted to generate simulated data, since this kind of model structure is quite representative of data collected from many real industrial processes, because the number of inner sources of variability that drives process behaviour is usually of a much smaller dimensionality than the number of measured variables (Burnham *et al.*, 1999; MacGregor & Kourti, 1998). The latent variable model employed has the following form:

$$\begin{aligned} X &= \mathbf{1}_n \cdot \mu_X{}^T + TP + E \\ Y &= \mathbf{1}_n \cdot \mu_Y{}^T + TQ + F \end{aligned} \tag{5.20}$$

where μ_X and μ_Y are the $m \times 1$ and $k \times 1$ vectors with the column averages of X and Y, $\mathbf{1}_n$ is a $n \times 1$ vector of ones, X is the $n \times m$ matrix of input data, Y is the $n \times k$ matrix of output data, T is the $n \times a$ matrix of latent variables that constitute the inner variability source, structuring both the input and output data matrices, E and F are $n \times m$ and $n \times k$ matrices of random errors, P and Q are $a \times m$ and $a \times k$ matrices of coefficients.

The model used in the simulations consists of five latent variables ($a = 5$) that follow a multivariate normal distribution with zero means and a diagonal covariance (I_a, i.e., the identity matrix of dimension a). The dimension of the input space is set equal to 10 and that of the output space equal to 1 ($m = 10,\ k = 1$). Rows of the P matrix form an a-orthonormal set of vectors with dimension m. The same applies to matrix Q, that consists of an a-orthonormal set of vectors with dimension k.

Each element of matrices E and F, of random errors, is drawn from a normal distribution with zero mean and standard deviation given by the uncertainty level associated with that specific variable (column of X or Y) for a particular observation (row). These uncertainties where allowed to vary, and this variation is characterized by

the "heterogeneity level" (HLEV), that measures the degree of variation or heterogeneity of uncertainties from observation to observation: HLEV=1 means a low variation of noise uncertainty or standard deviation from observation to observation, while HLEV=2 stands for a highly heteroscedastic behaviour for the noise uncertainties. More specifically, for variable X_i, uncertainties along the observation index are randomly generated from a uniform distribution centred at $\overline{u}\left(X_i\right)$ (average uncertainty for a given variable), with range given by $R(HLEV) = K_2(HLEV) \times \overline{u}\left(X_i\right)$, where K_2=0.01 (if HLEV=1; low heterogeneity level) or K_2=1 (if HLEV=2; high heterogeneity level), i.e.:

$$u\left(X_i(k)\right) \sim U\left[\overline{u}\left(X_i\right) - \frac{R(HLEV)}{2}, \overline{u}\left(X_i\right) + \frac{R(HLEV)}{2}\right] \tag{5.21}$$

In the present study, $\overline{u}\left(X_i\right)$ was kept constant at 0.5 times the theoretical standard deviation calculated for each noiseless variable.

5.2.1 Case Study 1: Complete Heteroscedastic Noise

With the goal of evaluating overall performance of the methods under different uncertainty structures for the measurements errors, the following sequence of steps was adopted:

i. Set the tuning parameters for each method and for each set of conditions (number of latent dimensions for PLS and PCR methods, and ridge parameter for RR methods). For PLS and PCR methods, $a = 5$. Regarding ridge methods, the ridge parameter was selected using cross-validation and the generation of a logarithmic grid in the range of plausible values (the criterion used in cross-validation is based upon the RMSEPW measure). This procedure is repeated 10 times, and the median of the best values is chosen as the tuning parameter to be used in the simulations. Variables are "auto-scaled" in all methods, except for OLS, MLS and MLMLS.

ii. For each scenario of HLEV (1 or 2), two noiseless data sets are generated according to the latent variable model presented above: a training or reference

noiseless data set and a test noiseless data set, both with 100 multivariate observations. Furthermore, a random sequence of uncertainties (noise standard deviations) for all the observations, belonging to each variable, is generated according to HLEV.

iii. Zero-mean Gaussian noise, with standard deviation given by the uncertainties calculated in ii., is generated and added to the noiseless training and testing data sets, after which a model is estimated according to each linear regression method (using the training data set) and its prediction performance evaluated (using the test data set). This process of noise addition, followed by parameter estimation and prediction, is repeated 100 times, and the corresponding performance metrics saved for future analysis.

Performance metrics used for prediction assessment are the square root of the weighted mean square error of prediction in the test set (RMSEPW), where the weights are the result of combining the predictor and response uncertainties, and the more familiar root mean square error of prediction (RMSEP):

$$RMSEPW(i) = \sqrt{\frac{1}{n}\sum_{k=1}^{n} \frac{\left(y(k) - \hat{y}(k)\right)^2}{uy(k)^2 + \left(uX(k,:)^{*2}\right)^T B^{*2}}} \ , \ i = 1,100 \tag{5.22}$$

$$RMSEP(i) = \sqrt{\frac{1}{n}\sum_{k=1}^{n} \left(y(k) - \hat{y}(k)\right)^2} \ , \ i = 1,100 \tag{5.23}$$

where n is the number of observations in the test set.

At the end of the simulations, we do have 100 values for the above metrics available for comparing the performances achieved by the different methods, under a given noise structure scenario. In order to take into account both the individual variability of the performance metrics for the different methods, as well as their mutual correlations, we based our comparison strategy in paired t-tests among all the different combinations of methods. Therefore, for each simulation scenario, paired t-tests were used to determine whether method A is better than method B (a “Win” for method A), performs worse (a

"Loose"), or if there is no statistical significant difference between both of methods A and B (a "tie"), for a given significance level (we used $\alpha = 0.01$). For the sake of simplicity, only the table with t-statistics and the corresponding plots with number of "Wins", "Looses" and "Ties" are presented here, for each simulation scenario studied.

Alternatively, multiple comparison methods (Kendall *et al.*, 1983; Scheffé, 1959) could also have been adopted, especially if one wants to have tight control over the overall significance level of the test performed. However, these types of methods are usually quite conservative, getting less sensitive to differences as the number of methods under comparison increases. For instance, a study where six methods were involved and significant differences apparently did exist, resulted in no difference being detected between any of the methods at a reasonable level of significance, using a Tukey's Test based multiple comparison approach (Indahl & Naes, 1998). Since we are comparing sixteen methods all together, the sensitivity of such a test would be even more affected, and therefore the choice went towards the adoption of an alternative, more sensitive approach. This comes at the cost of incurring in higher overall Type I errors rates than the significance level used for each method, but as long as this limitation is kept in mind, our results still provide a sound basis for establishing the kind of general guidelines we are interested in identifying.

Table 5.5 and Figure 5.1 present the comparison results obtained for scenario HLEV=1, using RMSEP as performance metric (since the trends for RMSEP and RMSEPW do not differ significantly, only those for the more familiar RMSEP are presented here).

Table 5.5. Results for the t-values obtained for the paired t-tests conducted to assess the statistical significance of the difference between RMSEP values obtained with method corresponding to line *i* and that for column *j*, i.e., RMSEP(method *i*) – RMSEP(method *j*) (* indicates a non-significant t value at $\alpha = 0.01$), using 100 replications under a simulation scenario with HLEV=1 (without missing data).

	PLS	uPLS1	uPLS2	uPLS3	uPLS4	uPLS5	RR	rMLS	rMLMLS	PCR	MLPCR	MLPCR1	MLPCR2	OLS	MLS	MLMLS
PLS	0	-11,106	-4,397	10,546	1,633*	11,292	-3,597	8,271	9,627	9,967	10,55	-3,57	10,801	-4,968	-11,907	5,53
uPLS1	11,106	0	9,463	15,529	11,03	15,636	10,989	13,22	14,216	16,003	16,117	9,302	16,079	10,917	-10,929	11,858
uPLS2	4,397	-9,463	0	11,228	4,475	11,959	4,543	8,564	9,468	12,243	12,478	0,923*	12,416	4,395	-11,639	6,134
uPLS3	-10,546	-15,529	-11,228	0	-10,564	-0,102*	-11,822	-7,799	-5,841	0,891*	1,807*	-11,539	2,306*	-11,911	-12,147	-10,166
uPLS4	-1,633*	-11,03	-4,475	10,564	0	10,6	-2,743	6,481	8,366	9,322	9,807	-4,012	10,244	-3,301	-11,933	3,719
uPLS5	-11,292	-15,636	-11,959	0,102*	-10,6	0	-12,856	-9,62	-7,596	1,051*	1,813*	-11,466	2,72	-12,922	-12,143	-11,299
RR	3,597	-10,989	-4,543	11,822	2,743	12,856	0	9,689	11,15	10,918	11,452	-3,507	12,007	-15,598	-11,844	7,214
rMLS	-8,271	-13,22	-8,564	7,799	-6,481	9,62	-9,689	0	7,789	7,712	8,004	-6,892	9,058	-9,844	-11,98	-7,104
rMLMLS	-9,627	-14,216	-9,468	5,841	-8,366	7,596	-11,15	-7,789	0	5,291	5,786	-7,805	6,888	-11,248	-12,074	-9,574
PCR	-9,967	-16,003	-12,243	-0,891*	-9,322	-1,051*	-10,918	-7,712	-5,291	0	1,744*	-13,521	2,559*	-11,003	-12,147	-8,88
MLPCR	-10,55	-16,117	-12,478	-1,807*	-9,807	-1,813*	-11,452	-8,004	-5,786	-1,744*	0	-14,096	1,935*	-11,533	-12,09	-9,499
MLPCR1	3,57	-9,302	-0,923*	11,539	4,012	11,466	3,507	6,892	7,805	13,521	14,096	0	13,698	3,364	-11,832	5,279
MLPCR2	-10,801	-16,079	-12,416	-2,306*	-10,244	-2,72	-12,007	-9,058	-6,888	-2,559*	-1,935*	-13,698	0	-12,081	-12,157	-10,185
OLS	4,968	-10,917	-4,395	11,911	3,301	12,922	15,598	9,844	11,248	11,003	11,533	-3,364	12,081	0	-11,84	7,604
MLS	11,907	10,929	11,639	12,147	11,933	12,143	11,844	11,98	12,074	12,147	12,09	11,832	12,157	11,84	0	11,972
MLMLS	-5,53	-11,858	-6,134	10,166	-3,719	11,299	-7,214	7,104	9,574	8,88	9,499	-5,279	10,185	-7,604	-11,972	0

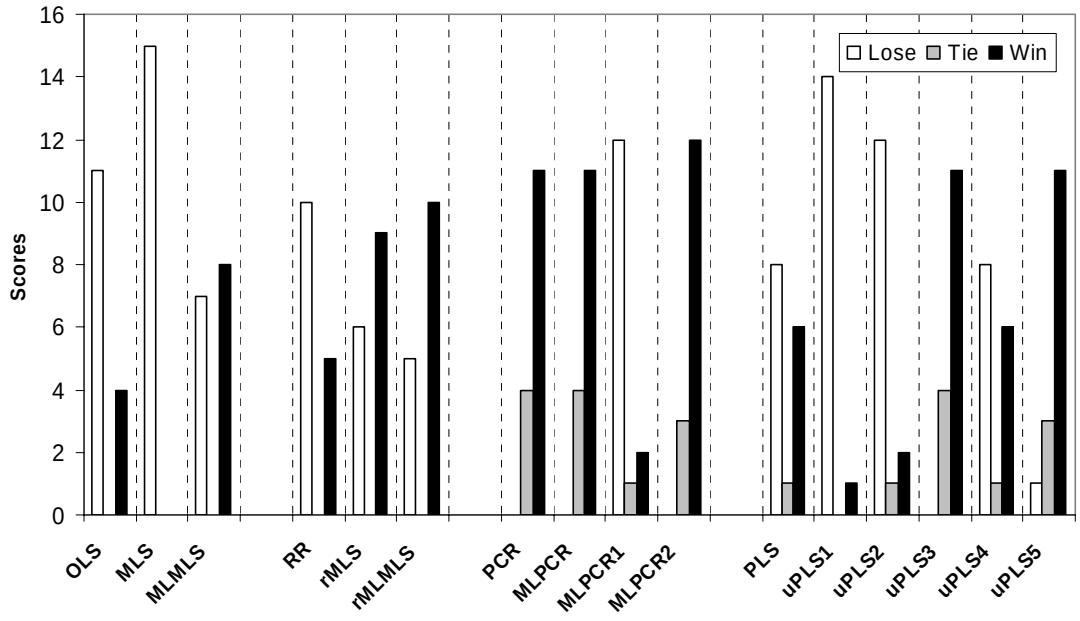


Figure 5.1. Number of "Looses", "Ties" and "Wins" for each method, under simulation scenario with HLEV=1 (using RMSEP).

Examining first the performance of methods belonging to the same group, it is possible to extract the following remarks from this simulation scenario:

- *OLS group.* MLS performs worse than OLS and MLMLS shows the best performance among the three methods. In general terms, comparing all the methods where MLS and MLMLS have similar roles (e.g. uPLS1/uPLS2, rMLS/rMLMLS, MLPCR1/MLPCR2), the second version never resulted in worse results and, as a matter of fact, almost always significantly improved them;
- *RR group*. Both rMLS and rMLMLS conducted to improved results relatively to those obtained by RR;
- *PCR group*. MLPCR does not improve over PCR predictive results, but MLPCR2 leads to an improvement;
- *PLS group*. Methods uPLS3 and uPLS5, both using uncertainty-based estimation of the relevant covariance matrices for PLS, present the best performance. Their similar performance results can be explained by the fact that, under mild

homoscedastic situations and if the variables present approximately equal uncertainties associated with their values, the orthogonal and non-orthogonal projections almost coincide. The same applies for the comparison of PLS and uPLS4, both using PLS weighting vectors but different projection strategies.

Comparing the results obtained for all the methods against each other, it is possible to note that MLPCR2 is the one that presented the best overall performance, followed by PCR, MLPCR, uPLS3 and uPLS5.

Figure 5.2 summarizes the results obtained for condition HLEV=2 (Table 5.6). A comparison of performances regarding methods within the PLS group shows that those methods that estimate the covariance matrices using uncertainty information (uPLS3, uPLS5) present better performance then their counterparts that use the same projection strategies (uPLS4, PLS, respectively). However, looking now to the methods that differ only on the projection methodology, it is possible to see that those that are based on orthogonal projections achieve better results that the ones based upon non-orthogonal maximum likelihood projections. This result is quite interesting, and will be further commented in the discussion section. In the PCR group it can observed that all methods perform quite well. As for the remaining groups of methods, the trends mentioned for HLEV=1 remain roughly valid. MLPCR2 continues to be the method with the best overall performance, followed by MLPCR and a group of methods that includes MLPCR1, PCR, and uPLS5.

Table 5.6. Results for the t-values obtained for the paired t-tests conducted to assess the statistical significance of the difference between RMSEP values obtained with method corresponding to line *i* and that for column *j*, i.e., RMSEP(method *i*) – RMSEP(method *j*) (* indicates a non-significant t value at $\alpha = 0.01$), using 100 replications under a simulation scenario with HLEV=2 (without missing data).

	PLS	uPLS1	uPLS2	uPLS3	uPLS4	uPLS5	RR	rMLS	rMLMLS	PCR	MLPCR	MLPCR1	MLPCR2	OLS	MLS	MLMLS
PLS	0	-11,063	-4,383	-0,977*	-11,394	13,343	-4,516	8,769	9,764	12,347	18,754	10,306	21,196	-5,599	-10,11	7,063
uPLS1	11,063	0	9,348	10,669	7,197	14,01	10,869	13,37	13,55	13,779	16,349	13,608	17,2	10,805	-9,602	12,624
uPLS2	4,383	-9,348	0	2,946	-3,951	10,486	3,965	9,957	10,108	11,809	16,817	11,09	19,194	3,831	-10,138	7,735
uPLS3	0,977*	-10,669	-2,946	0	-11,288	11,281	0,483*	5,433	6,096	10,379	17,466	11,033	19,43	0,337*	-10,108	4,195
uPLS4	11,394	-7,197	3,951	11,288	0	17,043	11,091	13,169	13,697	16,359	22,182	17,407	24,258	10,943	-9,959	12,283
uPLS5	-13,343	-14,01	-10,486	-11,281	-17,043	0	-13,335	-4,693	-4,909	-0,043*	10,33	2,386*	14,104	-13,384	-10,362	-6,898
RR	4,516	-10,869	-3,965	-0,483*	-11,091	13,335	0	9,417	10,337	12,42	18,885	10,631	21,465	-13,467	-10,102	7,985
rMLS	-8,769	-13,37	-9,957	-5,433	-13,169	4,693	-9,417	0	1,836*	4,852	11,898	4,864	16,878	-9,557	-9,889	-4,35
rMLMLS	-9,764	-13,55	-10,108	-6,096	-13,697	4,909	-10,337	-1,836*	0	4,646	12,715	4,575	18,665	-10,447	-10,114	-5,724
PCR	-12,347	-13,779	-11,809	-10,379	-16,359	0,043*	-12,42	-4,852	-4,646	0	13,239	2,526*	14,165	-12,467	-10,273	-6,366
MLPCR	-18,754	-16,349	-16,817	-17,466	-22,182	-10,33	-18,885	-11,898	-12,715	-13,239	0	-4,492	5,339	-18,892	-10,265	-13,014
MLPCR1	-10,306	-13,608	-11,09	-11,033	-17,407	-2,386*	-10,631	-4,864	-4,575	-2,526*	4,492	0	7,958	-10,745	-10,248	-6,649
MLPCR2	-21,196	-17,2	-19,194	-19,43	-24,258	-14,104	-21,465	-16,878	-18,665	-14,165	-5,339	-7,958	0	-21,468	-10,568	-17,806
OLS	5,599	-10,805	-3,831	-0,337*	-10,943	13,384	13,467	9,557	10,447	12,467	18,892	10,745	21,468	0	-10,098	8,199
MLS	10,11	9,602	10,138	10,108	9,959	10,362	10,102	9,889	10,114	10,273	10,265	10,248	10,568	10,098	0	9,866
MLMLS	-7,063	-12,624	-7,735	-4,195	-12,283	6,898	-7,985	4,35	5,724	6,366	13,014	6,649	17,806	-8,199	-9,866	0

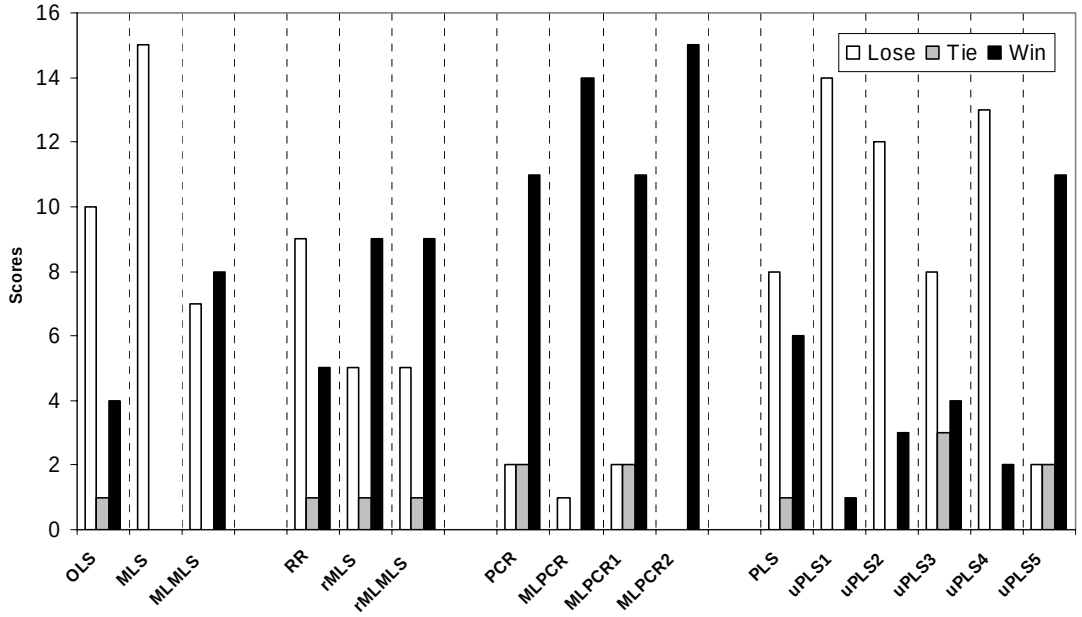


Figure 5.2. Number of "Looses", "Ties" and "Wins" for each method, under simulation scenario with HLEV=2 (using RMSEP).

5.2.2 Case Study 2: Handling Missing Data

In this second case study, the prediction performance of the several methods is analysed when a fraction of data is missing (both in the model estimation and prediction stages), and a very simple strategy for handling missing data is adopted: *mean substitution*. For uncertainty based methods, one also has to specify the uncertainty values associated with these inputted values, for which the standard deviations of the respective variables during normal operation were adopted. Other more sophisticated methodologies for missing data imputation during model estimation are also available for regression methods, specially PLS and PCR (Walczak & Massart, 2001), as well as methods for handling missing data once we have already an estimated model at our disposal (Nelson *et al.*, 1996). Analogous approaches can also be developed for the uncertainty based techniques, that only require the estimated values to fill in existing blanks and their associated uncertainties. However, the aim of this study is to assess the extent to which one can easily handle missing data in model estimation and prediction (i.e., with minimum assumptions regarding missing values and with the least modification over standard procedures), taking advantage of the possibility of using uncertainty information. That being the case, it was decided to keep the same replacement strategy

amongst all the tested methods, so that the real advantage of handling such an additional piece of information, provided by measurement uncertainties, can be easily evaluated and compared with the current alternatives.

As our focus here is related with an evaluation of the methods regarding prediction when missing data is present, we adopted a simulation structure which is now different from that of case study 1. For *each simulation*, the following steps are repeated and the corresponding results saved:

i. Generate a new latent variable model (matrices Q and P) and noiseless data to be used for model estimation and prediction assessment. Generate also measurement uncertainties to be associated with each non-missing value, according to the value of $HLEV$ used in each simulation study;

ii. Generate a new "missing data mask" that removes (on average) a chosen percentage of the data matrix [X|Y]. We used a target percentage of 20%, both for the reference and test data sets;

iii. Generate and add noise to the noiseless data that were not removed, according to the measurement uncertainties generated in i.;

iv. Replace missing data with column means for the data set used to estimate the model, and calculate the associated uncertainties using the columns standard deviations, for the same data set;

v. Estimate models using the data set constructed in iv.;

vi. For the test data set, do the same operation as in iv. (using the same values for the input values and uncertainties) and calculate the predicted value for the output variable. Calculate overall performance metrics (RMSEPW and RMSEP).

The results obtained with HLEV=1 are presented in Table 5.7 and Figure 5.3, where it can be seen that within the PLS group methods uPLS5 and uPLS3 lead to improved predictive performances, but now with uPLS3 presenting better results that uPLS5, i.e., the non-orthogonal projection seems to bring some added value when missing data is present, under homoscedastic scenarios. In the PCR group, all MLPCR methods outperform conventional PCR. As for the other groups, results obtained follow the same

trends verified when no missing data were present. In global terms, MLPCR2 presents the best overall performance, followed by MLPCR1, MLPCR and uPLS3.

Table 5.7. Results for the t-values obtained for the paired t-tests conducted to assess the statistical significance of the difference between RMSEP values obtained with method corresponding to line *i* and that for column *j*, i.e., RMSEP(method *i*) – RMSEP(method *j*) (* indicates a non-significant t value at $\alpha = 0.01$), using 100 replications under a simulation scenario with HLEV=1 (with 20% of missing data).

	PLS	uPLS1	uPLS2	uPLS3	uPLS4	uPLS5	RR	rMLS	rMLMLS	PCR	MLPCR	MLPCR1	MLPCR2	OLS	MLS	MLMLS
PLS	0	-9,107	0,853*	12,039	2,681	8,244	-1,618*	10,728	12,484	8,437	20,856	12,416	23,111	-2,042*	-12,128	10,653
uPLS1	9,107	0	9,588	15,103	9,488	12,035	9,032	13,484	13,595	12,337	17,258	15,052	20,786	9,009	-11,689	12,79
uPLS2	-0,853*	-9,588	0	10,392	1,288*	5,611	-0,954*	9,939	9,994	6,305	16,561	12,589	25,237	-0,983*	-12,203	8,761
uPLS3	-12,039	-15,103	-10,392	0	-8,71	-9,281	-12,126	-4,146	-4,214	-7,42	7,129	7,271	17,345	-12,133	-12,42	-3,703
uPLS4	-2,681	-9,488	-1,288*	8,71	0	2,097*	-2,798	4,474	4,769	2,59*	12,169	10,532	17,821	-2,836	-12,08	4,873
uPLS5	-8,244	-12,035	-5,611	9,281	-2,097*	0	-8,276	4,498	6,584	1,945*	19,405	10,571	24,402	-8,266	-12,311	4,402
RR	1,618*	-9,032	0,954*	12,126	2,798	8,276	0	10,835	12,514	8,493	20,957	12,462	23,148	-4,416	-12,125	10,786
rMLS	-10,728	-13,484	-9,939	4,146	-4,474	-4,498	-10,835	0	0,747*	-2,983	10,428	9,606	22,91	-10,854	-12,363	0,689*
rMLMLS	-12,484	-13,595	-9,994	4,214	-4,769	-6,584	-12,514	-0,747*	0	-3,825	11,306	9,118	22,558	-12,508	-12,367	-0,055*
PCR	-8,437	-12,337	-6,305	7,42	-2,59*	-1,945*	-8,493	2,983	3,825	0	20,771	10,26	23,838	-8,486	-12,349	2,864
MLPCR	-20,856	-17,258	-16,561	-7,129	-12,169	-19,405	-20,957	-10,428	-11,306	-20,771	0	3,194	11,822	-20,911	-12,606	-9,264
MLPCR1	-12,416	-15,052	-12,589	-7,271	-10,532	-10,571	-12,462	-9,606	-9,118	-10,26	-3,194	0	4,623	-12,477	-12,581	-9,655
MLPCR2	-23,111	-20,786	-25,237	-17,345	-17,821	-24,402	-23,148	-22,91	-22,558	-23,838	-11,822	-4,623	0	-23,15	-12,843	-21,715
OLS	2,042*	-9,009	0,983*	12,133	2,836	8,266	4,416	10,854	12,508	8,486	20,911	12,477	23,15	0	-12,123	10,816
MLS	12,128	11,689	12,203	12,42	12,08	12,311	12,125	12,363	12,367	12,349	12,606	12,581	12,843	12,123	0	12,521
MLMLS	-10,653	-12,79	-8,761	3,703	-4,873	-4,402	-10,786	-0,689*	0,055*	-2,864	9,264	9,655	21,715	-10,816	-12,521	0

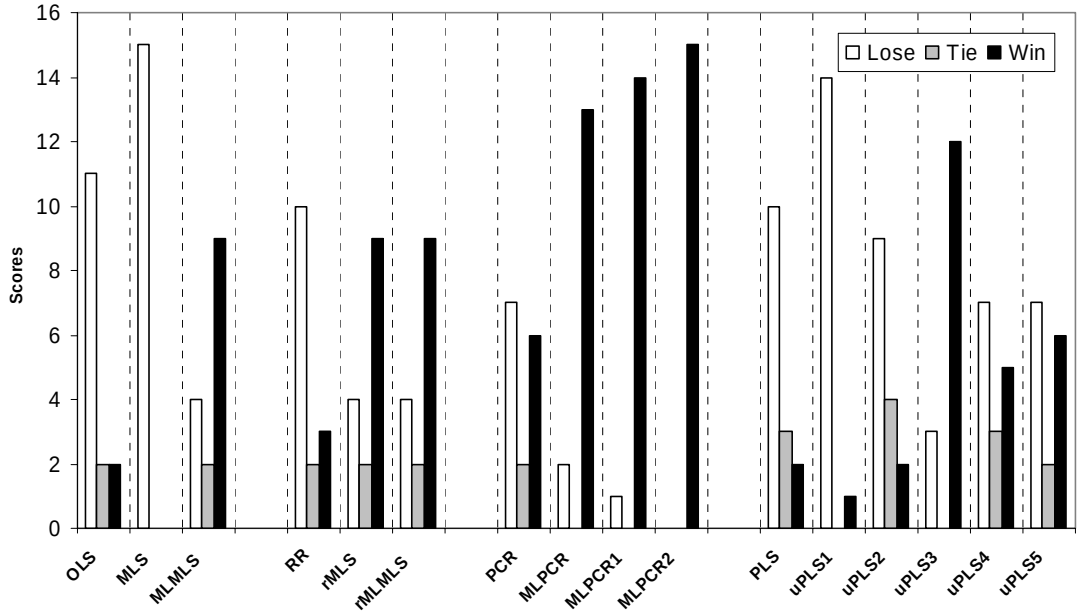


Figure 5.3. Number of "Looses", "Ties" and "Wins" for each method, under simulation scenario with HLEV=1 and 20% of missing data (using RMSEP).

Analysing now the results for HLEV=2 (Table 5.8 and Figure 5.4), it is also possible to verify that uPLS3 and uPLS5 still show the best predictive performance within the PLS

group, but now with uPLS3 presenting lower scores relatively to the previous scenario (HLEV=1), a result that is consistent with what was verified in case study 1. In the global comparison, after MLPCR2 we can find methods MLPCR1 and MLPCR.

Therefore, under the conditions adopted for this simulation study, it can be concluded that MLPCR methods tend to have the best overall performance in the presence of missing data.

Table 5.8. Results for the t-values obtained for the paired t-tests conducted to assess the statistical significance of the difference between RMSEP values obtained with method corresponding to line *i* and that for column *j*, i.e., RMSEP(method *i*) – RMSEP(method *j*) (* indicates a non-significant t value at $\alpha = 0.01$), using 100 replications under a simulation scenario with HLEV=2 (with 20% of missing data).

	PLS	uPLS1	uPLS2	uPLS3	uPLS4	uPLS5	RR	rMLS	rMLMLS	PCR	MLPCR	MLPCR1	MLPCR2	OLS	MLS	MLMLS
PLS	0	-11,458	-7,485	5,177	-1,256*	4,559	-1,923*	9,239	10,813	5,688	21,39	14,706	25,185	-2,336*	-12,401	9,747
uPLS1	11,458	0	7,207	14,849	9,666	12,836	11,444	14,699	15,601	13,512	22,982	21,503	27,794	11,426	-11,874	14,843
uPLS2	7,485	-7,207	0	11,833	4,938	10,239	7,474	13,792	15,253	11,044	24,851	20,846	32,192	7,447	-12,362	13,787
uPLS3	-5,177	-14,849	-11,833	0	-6,065	-2,771	-5,281	0,733*	1,342*	-1,927*	13,954	13,773	25,351	-5,292	-12,457	1,13*
uPLS4	1,256*	-9,666	-4,938	6,065	0	3,328	1,175*	6,482	6,944	3,989	15,797	15,269	22,57	1,158*	-12,474	6,923
uPLS5	-4,559	-12,836	-10,239	2,771	-3,328	0	-4,708	3,79	6,157	0,848*	20,199	14,027	27,23	-4,712	-12,135	4,61
RR	1,923*	-11,444	-7,474	5,281	-1,175*	4,708	0	9,495	11,118	5,913	21,718	14,868	25,552	-3,926	-12,398	10,002
rMLS	-9,239	-14,699	-13,792	-0,733*	-6,482	-3,79	-9,495	0	2,956	-3,782	12,673	12,837	24,41	-9,512	-12,734	1,849*
rMLMLS	-10,813	-15,601	-15,253	-1,342*	-6,944	-6,157	-11,118	-2,956	0	-5,535	12,82	12,539	25,805	-11,117	-12,37	-1,26*
PCR	-5,688	-13,512	-11,044	1,927*	-3,989	-0,848*	-5,913	3,782	5,535	0	22,047	14,569	28,148	-5,915	-12,431	4,114
MLPCR	-21,39	-22,982	-24,851	-13,954	-15,797	-20,199	-21,718	-12,673	-12,82	-22,047	0	2,112*	11,978	-21,665	-13,114	-11,605
MLPCR1	-14,706	-21,503	-20,846	-13,773	-15,269	-14,027	-14,868	-12,837	-12,539	-14,569	-2,112*	0	8,448	-14,876	-13,204	-12,396
MLPCR2	-25,185	-27,794	-32,192	-25,351	-22,57	-27,23	-25,552	-24,41	-25,805	-28,148	-11,978	-8,448	0	-25,531	-13,416	-23,146
OLS	2,336*	-11,426	-7,447	5,292	-1,158*	4,712	3,926	9,512	11,117	5,915	21,665	14,876	25,531	0	-12,397	10,02
MLS	12,401	11,874	12,362	12,457	12,474	12,135	12,398	12,734	12,37	12,431	13,114	13,204	13,416	12,397	0	12,537
MLMLS	-9,747	-14,843	-13,787	-1,13*	-6,923	-4,61	-10,002	-1,849*	1,26*	-4,114	11,605	12,396	23,146	-10,02	-12,537	0

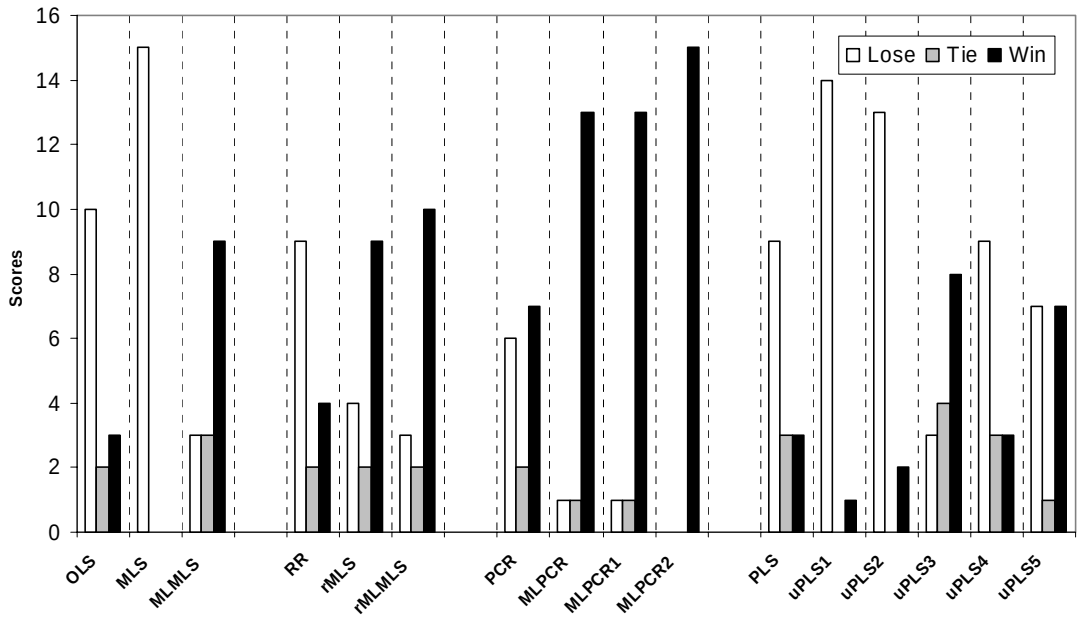


Figure 5.4. Number of "Looses", "Ties" and "Wins" for each method, under simulation scenario with HLEV=2 and 20% of missing data (using RMSEP).

It is important to point out that when adopting a methodology that integrates data uncertainty one follows the same calculation procedure adopted for the situation where no data is missing, simply replacing the missing elements with rough estimates that will be properly weighted by the algorithms, according to their associated uncertainties. However, if there are better estimates available, for instance arising from more sophisticated imputation techniques, one can also integrate them as well, without any further changes, as long as these are provided along with their associated uncertainties.

5.3 Discussion

Results presented in the previous section highlight not only the potential of using all the information that is available (data and associated uncertainties), but also the difficulty that such a task may encompass regarding model estimation. In fact, there were some unexpected results and relevant issues have been identified and deserve being discussed here.

First of all, we would like to stress that, even though simulation results are strictly valid within the conditions established, they can provide useful guidelines for real processes that present structural similarities with them. Then, it is also important to bear in mind that the fact classical methods do not make explicit use of uncertainty information may not be very relevant if it represents just a small part of the global variability exhibited by variables. These are however tacit assumptions, made by conventional approaches, quite often not verified or clearly stated, and the main purposes of the work presented in this chapter were precisely to bring the issue of data uncertainty into the priorities for the analyst, who should explicitly address it in a preliminary stage of data analysis, as well as to present, develop and test procedures that do exploit and take advantage of data uncertainty information. In general, uncertainty-based methods presented here are only expected to bring potentially more added value under contexts where uncertainty is quite high (noisy environments) or experiment large variations. In other words, these methods should complement their classical counterparts, depending on the noise characteristics that prevail in measured data.

Still regarding model estimation, some convergence problems were found in MLMLS, something that is not unusual in approaches based upon numerical optimization of a non-linear objective function. However, problems in MLPCR2 due to the non convergence of MLMLS were found to be quite rare. From the experience that we have gathered so far, no limitations were found regarding the implementation of MLPCR2 in the analysis of real industrial data.

The poor performance of MLS under the scenarios considered here, where predictors are strongly correlated, may indicate that the inversion operation undertaken at each iteration is interfering with its performance (the matrix to be inverted in this method becomes quite ill-conditioned under collinear situations of the predictors). Results obtained for the ridge regularization of MLS (rMLS) show an effective stabilization of this operation.

As for PLS methods, the extensive solution of small optimization problems can make uPLS1 and uPLS2 more prone to numerical convergence problems than the original PLS method, something that does not occur with the remaining uncertainty-based PLS methods (uPLS3, uPLS4 and uPLS5), since they are based upon the estimation of covariance matrices and projection operations. In spite of the fact that uPLS1 and uPLS2 represent efforts towards the explicit integration of uncertainty information into the algorithmic structure of PLS, some simplifications were introduced into it. Namely, the uncertainty of loading vectors and weights was neglected. Future developments should consider these issues, with the same concerns applying also to MLPCR methods, where uncertainty in the loadings is also neglected when the propagation of uncertainties to the scores is carried out. The assumed independence of uncertainties in the scores for the regression step in MLPCR1 and MLPCR2 may also deserve more attention in future studies. The relatively poor performance of uPLS1 under the simulations conditions studied here, where a latent variable model is adopted for generating data, in comparison to the good results obtained under situations where a multivariate linear regression model was adopted (with regressors having several degrees of correlation), as presented elsewhere (Reis & Saraiva, 2004b), is worthwhile noticing. It is therefore advisable to, whenever possible, take advantage of current availability of computation power and conduct exploratory simulation studies under conditions close to the intended application (e.g., similar predictors' correlations, noise

structures), in order to get more insight into the problem, as well as regarding the most adequate approaches to handle it.

When comparing, under heteroscedastic situations (Figure 5.2), PLS methods that adopt the same estimation procedure for the covariance matrices but differ in the projection phase (as happens with pairs PLS/uPLS4, uPLS3/uPLS5), one can see that the use of uncertainty-based maximum-likelihood non-orthogonal projections seems to be detrimental for prediction relatively to orthogonal projections. In fact, a separate simulation study showed evidence towards a reduced variance of the orthogonal projection scores, when compared to the one exhibited by maximum likelihood projection scores. Apparently, for heteroscedastic scenarios, oscillations in the non-orthogonal projection line may also bring some added variability to the scores, other than the one strictly arising from variability due to noise sources. This increased dispersion in the reduced space of the scores, usually the one relevant for prediction purposes, can increase prediction uncertainty due to poorly estimated models, something that is in line with the results presented in Figure 5.2.

Finally, although we have focused here on steady-state applications, the above mentioned approaches can also be used under the context of dynamic models, namely through the consideration of lagged variables (Ku *et al.*, 1995; Ricker, 1988; Shi & MacGregor, 2000).[26]

5.4 Conclusions

In this chapter the importance of specifying measurement uncertainties, and how this information can be used in the estimation of linear regression models, was addressed. Under the conditions covered in this study, method MLPCR2 presented the best overall predictive performance. In general, those methods based on MLMLS present improvements over their counterparts based on MLS.

[26] However, PLS methods based upon uncertainty-based estimation of covariance matrices do need some modifications in order to cope with noise correlations appearing now, with the use of lagged variables.

Several real world applications are associated with contexts where uncertainty-based methods can be used with potential benefits. These methods can also be applied with added value to the analysis of the approximation coefficients for a given selected scale, arising from MRD frameworks (Chapter 4).

Chapter 6. Integrating data uncertainty information in process optimization

In this chapter, a complementary situation regarding the use of data uncertainties, is addressed. In particular, we are now concerned with its application when a process model is assumedly available, as well as information regarding measurement and actuation uncertainties, and the main purpose is to derive an optimal operation policy, in the sense of achieving a certain, pre-defined, production goal.

6.1 Problem Formulation

The problem can be formulated as one where one aims to find optimal settings regarding a manipulated variables vector (Z), given a certain objective function (e.g., maximize some profit metric or minimize a cost function), for a given measurement of the vector of load variables (L). However, due to the presence of uncertainties, the following relevant issues do arise:

- *Measured* quantities (i.e., the loads, $\tilde{L}$, and the outputs, $\tilde{Y}$) are affected by measurement noise, with statistical characteristics defined by their associated uncertainty

$$\begin{aligned} \tilde{L} &= L + \varepsilon_L \\ \tilde{Y} &= Y + \varepsilon_Y \end{aligned} \tag{6.1}$$

with quantities marked with "~" being the values actually available, while L and Y are the corresponding true, but unknown, values for these quantities (Figure 6.1).

- Similarly, the set-point that we specify for the manipulated variables ($\tilde{Z}$) does not correspond to the exact true value of the manipulation action over the process. In fact, due to *actuation noise*, there is also here another source of uncertainty to be taken into account.

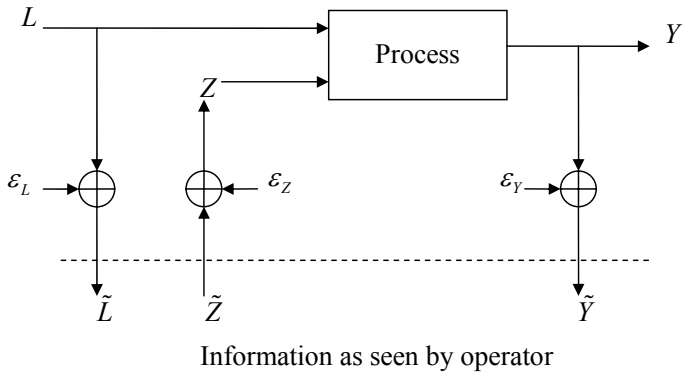


Figure 6.1. Schematic representation of measured quantities (as seen by an external operator and marked with "~") and the quantities that are actually interacting with the process.

Considering that our goal is to drive the process in such a way as to minimize some relevant cost function, $\phi(\cdot)$, the following formulation is proposed, incorporating measurement and actuation uncertainties in the calculation of the adequate values for the manipulated variables to be specified externally, when a given measurement for the load is acquired ($\tilde{L}$). As often happens in the formulation of optimization problems under uncertainty, the objective function comprises an expected value for the performance metric, taken over the space of uncertain parameters:

Formulation I

$$\begin{aligned} \underset{\tilde{Z}}{Min} \quad & E_{\Theta}\left\{\phi\left(L, Z, \tilde{Y}\right)\right\} \\ s.t. \quad & g\left(Y, L, Z\right) = 0 \\ & L = \tilde{L} - \varepsilon_L \\ & Z = \tilde{Z} + \varepsilon_Z \\ & \tilde{Y} = Y + \varepsilon_Y \end{aligned} \tag{6.2}$$

where $E_{\Theta}\{\cdot\}$ is the expectation operator

$$E_{\Theta}\{\phi\} = \int_{\Theta} \phi(\theta) j(\theta) d\theta \tag{6.3}$$

with $\theta = [\varepsilon_L^T, \varepsilon_Z^T, \varepsilon_Y^T]^T$ and $j(\theta)$ providing the joint probability density function for the uncertain quantities θ. The available model is represented by $g\left(Y, L, Z\right) = 0$, and we will assume here that the uncertainty associated with its parameters is negligible.[27]

In *Formulation I*, it is assumed that the relevant quantities for evaluation of the performance metric are the values of *L* and *Z* that really affect the process, as well as the *measured* value of the output. It is important to point out that these assumptions do not necessarily hold for all situations. For instance, sometimes the performance metric should be calculated with the "true" value of the output, *Y*, instead of $\tilde{Y}$ (*Formulation II*, see below), as is the case when output measurements become available with much less uncertainty in a subsequent stage (e.g., from off-line laboratory tests). Other times, only measured values should be used, because no better measurements or reconciliation

[27] If not, such uncertainties can also be incorporated in the problem formulation (Rooney & Biegler, 2001).

procedures can be adopted. The correct formulation is therefore case dependent, and should be tailored to each particular situation.

Formulation 2

$$\begin{aligned} \underset{\tilde{Z}}{Min} \quad & E_{\Theta}\left\{\phi\left(L,Z,Y\right)\right\} \\ s.t. \quad & g\left(Y,L,Z\right)=0 \\ & L=\tilde{L}-\varepsilon_{L} \\ & Z=\tilde{Z}+\varepsilon_{Z} \end{aligned} \tag{6.4}$$

In the example presented in this chapter, we also report the results obtained for the situation where uncertainties are not taken into account at all, and thus where the manipulated variable values are found by solving the following problem:

Formulation 3

$$\begin{aligned} \underset{\tilde{Z}}{Min} \quad & \phi\left(\tilde{L},\tilde{Z},\tilde{Y}\right) \\ s.t. \quad & g\left(\tilde{Y},\tilde{L},\tilde{Z}\right)=0 \end{aligned} \tag{6.5}$$

6.2 Illustrative Example

This example illustrates the integration of measurement uncertainties in process optimization decision-making. As referred before, the problem being addressed consists of calculating the values for the manipulated variables to be specified ($\tilde{Z}$) in order to minimize a cost function, when measurements for the loads become available ($\tilde{L}$). This particular case study is based on the following model, developed for a batch paper pulp pilot digester (Carvalho *et al.*, 2003):

$$\mathrm{TY}=55.2-0.39\times\mathrm{EA}+324/\left(\mathrm{EA}\times\log_{10}\mathrm{S}\right)-92.8\times\log_{10}(\mathrm{H})/\left(\mathrm{EA}\times\log_{10}\mathrm{S}\right) \tag{6.6}$$

This model relates pulp total yield (TY, %) with effective alkali (EA, a measure for the joint concentration of Na_2OH and Na_2S, the active elements in the cooking liquor, %), sulfidity (S, the percentage of Na_2S in the cooking liquor, %) and H factor (H, a function of the temperature profile across the batch).

Let us consider the situation where a cost function (L) penalizes deviations from a target value for TY (52%): penalty for lower values is due to fibre loss and that for higher values due to deterioration in other pulp properties. The cost function also takes into account the costs of S and H (proportional to their respective magnitudes):

$$L = \begin{cases} 100\left(\frac{TY_{sp}}{100} - \frac{TY}{100}\right) + \frac{S}{4} + \frac{H}{500} & \Leftarrow \quad TY \le TY_{sp} \\ 75^2\left(\frac{TY_{sp}}{100} - \frac{TY}{100}\right)^2 + \frac{S}{4} + \frac{H}{500} & \Leftarrow \quad TY > TY_{sp} \end{cases} \tag{6.7}$$

As an example, Figure 6.2 illustrates the shape of the assumed cost function for $S = 20$ and $H = 1000$.

In this example, EA is assumed to be a load variable, and thus our optimization goal consists of calculating the S and H values that minimize expected cost in the presence of uncertainties for both measurements and process actuations. *Formulations I*, *II* and *III* hold for this example, with $L = \text{EA}$, $Z = [\text{S H}]$ and $Y = \text{TY}$ (Table 6.1).

Table 6.1. Optimization formulations I, II and III, as applied to the present example.

Formulation I	Formulation II	Formulation III
$\underset{\tilde{S},\tilde{H}}{Min} \quad E_\Theta\left\{\phi\left(EA,S,H,\tilde{TY}\right)\right\}$	$\underset{\tilde{S},\tilde{H}}{Min} \quad E_\Theta\left\{\phi\left(EA,S,H,TY\right)\right\}$	$\underset{\tilde{S},\tilde{H}}{Min} \quad \phi\left(\tilde{EA},\tilde{S},\tilde{H},\tilde{TY}\right)$
$s.t. \quad g\left(TY,EA,S,H\right)=0$	$s.t. \quad g\left(TY,EA,S,H\right)=0$	$s.t. \quad g\left(\tilde{TY},\tilde{EA},\tilde{S},\tilde{H}\right)=0$
$EA = \tilde{EA} - \varepsilon_{EA}$	$EA = \tilde{EA} - \varepsilon_{EA}$	
$S = \tilde{S} + \varepsilon_S$	$S = \tilde{S} + \varepsilon_S$	
$H = \tilde{H} + \varepsilon_H$	$H = \tilde{H} + \varepsilon_H$	
$\tilde{TY} = TY + \varepsilon_{TY}$		

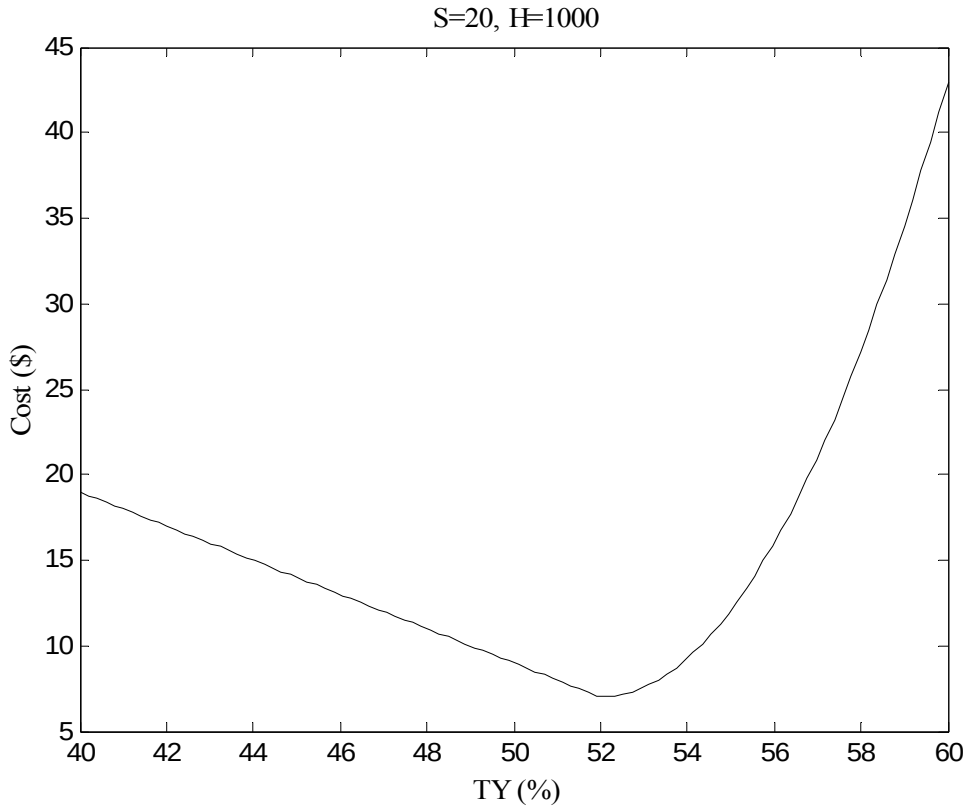


Figure 6.2. Cost function for deviations of TY from its target value (52%), for S=20 and H=1000.

We further assume that the vector of uncertain quantities, $\theta = [\varepsilon_{EA}, \varepsilon_S, \varepsilon_H, \varepsilon_{TY}]^T$, follows a multivariate normal distribution with zero mean and diagonal covariance given by:

$$\Sigma_\Theta = diag([2^2 \quad 2^2 \quad 50^2 \quad 4^2]) \tag{6.8}$$

where *diag* stands for the operator that converts a vector into a diagonal matrix with its elements along the main diagonal.

To illustrate the implementation of the formulations referred above, let us consider that the observed value for EA is 15% ($\tilde{EA} = 15$). Table 6.2 summarizes the results obtained for the manipulated variables ($\tilde{S}$ and $\tilde{H}$) and the average cost obtained with the objective function assumed under formulations I and II, with a third degree specialized cubature being used for estimation of expected values (Bernardo *et al.*, 1999).

Table 6.2. Solutions obtained under formulations I, II and III, and their associated average costs.

Solutions:	Average cost (formulation I) *($)*	Average cost (formulation II) *($)*
I. $\tilde{S}$=7.16 $\tilde{H}$=1602.0	10.80	5.93
II. $\tilde{S}$=7.83 $\tilde{H}$=1184.2	11.16	5.40
III. $\tilde{S}$=5.38 $\tilde{H}$=1274.6	25.46	8.17

From Table 6.2 we can see that under the simulation conditions considered here, and assuming that the relevant objective function is the one associated with formulation I, the optimal solution obtained when one disregards measurement and actuation uncertainties (formulation III) corresponds to an average cost increased by 136%. If the relevant objective function were the one corresponding to problem formulation II, the average cost increase would be 51%. It should also be noticed that the location of the optimal solution in the $(\tilde{S},\tilde{H})$ decision space, found if one ignores uncertainties, is quite distant from the optimal one.

The cost associated with the non consideration of these types of uncertainties decreases when their magnitude gets smaller. Figure 6.3 presents the results obtained for the three alternative problem formulations, when the covariance matrix for uncertain quantities is multiplied by a monotonically decreasing shrinkage factor, 0.9^i. As expected, the differences arising from the solutions associated with such three optimization formulations tend to vanish when measurement and actuation uncertainties decrease. Furthermore, average cost also decreases, because of the improved quality of information obtained from measurement devices and the better performance of final control elements, as one moves across the several simulation scenarios considered here.

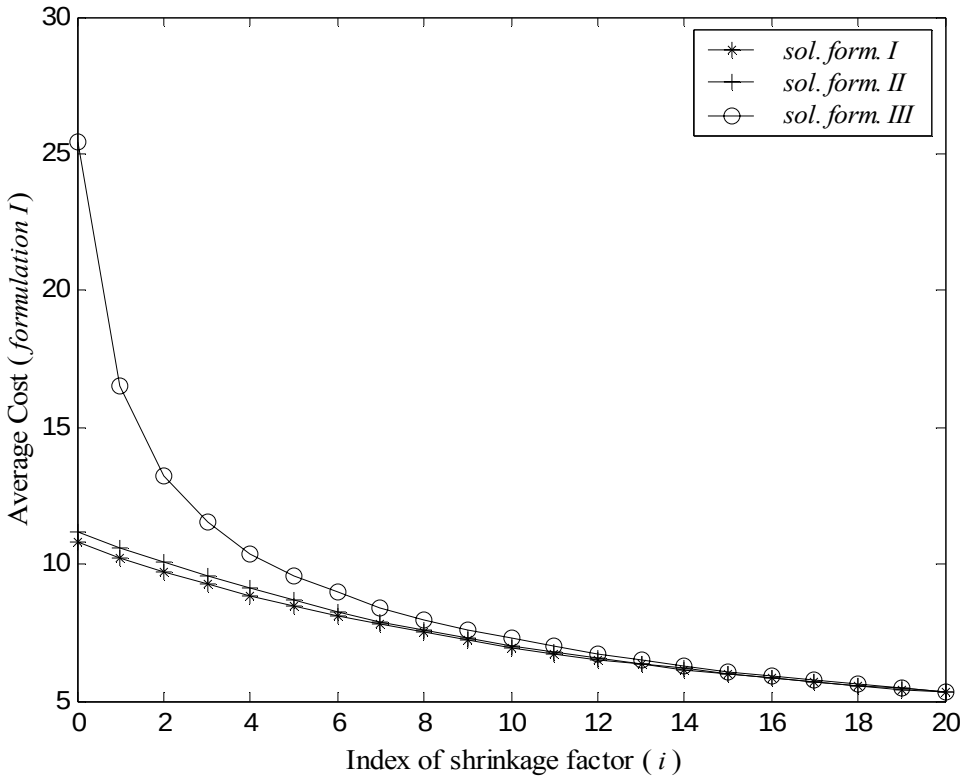


Figure 6.3. Behaviour of average cost (formulation I), corresponding to solutions for the three alternative problem formulations, using $\Sigma_{\Theta} \cdot 0.9^{i}$.

6.3 Conclusions

In this chapter we have addressed the integration of measurement and actuation uncertainties in process optimization problems. Several possible formulations were proposed, and the study carried out points the relevance of not neglecting measurement/manipulation uncertainties when addressing both on-line and off-line process optimization. In fact, the cost of performing process optimization under such circumstances (i.e., not considering uncertainty information) can lead to a significant increase in the cost function, in a situation such as the one illustrated in the example presented.

Chapter 7. Integrating Data Uncertainty Information in Multivariate Statistical Process Control

Current SPC methodologies (Section 3.1), based upon latent variables, do not take explicitly into consideration information regarding measurement uncertainty. As such, they do not directly explore this valuable piece of knowledge that, furthermore, is becoming increasingly available, given the recent developments on instrumentation technology and metrology. In this chapter we present a single-scale approach for conducting multivariate statistical process control (MSPC) that incorporates data uncertainty information. It is specially suited to perform process monitoring under noisy environments, i.e., when the signal to noise ratio is low, and, furthermore, the noise standard deviation (uncertainty), affecting each collected value, can vary over time, and is assumingly known.

Our approach is based upon a latent variable model structure, HLV (standing for heteroscedastic latent variable model), that explicitly integrates information regarding data uncertainty. Moderate amounts of missing data can also be handled in a coherent and fully integrated way through the HLV model. Several examples illustrate the added value achieved under noisy conditions by adopting such an approach, and an additional case study illustrates its application to a real industrial context, regarding pulp and paper product quality data analysis.

The statistical model based upon which the approach for integrating data uncertainties was built is presented in the next section. Then, a discussion about its relationship with other latent variable models is provided. The description of the proposed MSPC procedure, based on latent variables when measurements have heteroscedastic Gaussian behaviour, can be found in the third section, where we also show how the proposed approach can easily handle the presence of missing data. In the following section, several examples are presented in order to illustrate the various features of the methodology, including a continuation of the case study initiated in Section 4.4.4., based upon real industrial quality data collected from a pulp and paper mill.

7.1 Underlying Statistical Model

The underlying statistical model adopted[28] addresses the fairly common situation where a large number of measurements are being collected and stored, arising from many different devices and sources within an industrial process, that carry important pieces of information about the current state of operation. Quite often the underlying process phenomena, along with existing process constraints, induce a significantly smaller dimensionality being needed to adequately describe collected industrial data, than that given by the entire set of measured variables. In fact, for monitoring purposes, we are only interested in following what happens around the subspace where the overall normal process variability is concentrated. In this context, latent variable models do provide useful frameworks for modelling the relationships linking the whole set of measurements, arising from different sources, in terms of a fewer inner variability sources (Burnham *et al.*, 1999).

Therefore, let us consider the following latent variable multivariate linear relationship:

$$x(k) = \mu_X + A \cdot l(k) + \varepsilon_m(k) \tag{7.1}$$

[28] Parts of this model were already briefly introduced in Section 5.1.4, but we present the whole model in the present chapter in order to facilitate the exposition and to make it comprehensive and self-contained.

where x is the $n \times 1$ vector of measurements, μ_X is the $n \times 1$ mean vector of x, A is the $n \times p$ matrix of model coefficients, l is the $p \times 1$ vector of latent variables and ε_m is the $n \times 1$ vector of measurement noise. This model is completed by specifying the probability density functions associated with each random component:

$$\begin{aligned} & l(k) \sim iid\ N_p\left(\mathbf{0}, \Delta_l\right) \\ & \varepsilon_m(k) \sim id\ N_n\left(\mathbf{0}, \Delta_m(k)\right) \\ & l(k) \text{ and } \varepsilon_m(j) \text{ are independent } \forall k, j \end{aligned} \tag{7.2}$$

where N_p stands for the p-dimensional multivariate normal distribution, Δ_l is the covariance matrix for the latent variables (l), $\Delta_m(k)$ is the covariance matrix of the measurement noise at time k ($\varepsilon_m(k)$), given by $\Delta_m(k) = diag(\sigma_m^2(k))$ ($diag(u)$, represents a diagonal matrix with the elements of vector u along the main diagonal and $\sigma_m^2(k)$ is the vector of error variances for all the measurements at time k), $\mathbf{0}$ is an array of appropriate dimension, with only zeros in its entries. Thus, equations (7.1) and (7.2) basically consider that the multivariate variability of x can be adequately described by the underlying behaviour of a smaller number of p latent variables, plus noise added in the full variable space. We can also see that such a model essentially consists of two parts: one that captures the variability due to normal process sources ($\mu_X + A \cdot l(k)$), and the other that explicitly describes the characteristics of measurement noise or uncertainties ($\varepsilon_m(k)$), each of them having its own independent randomness. In the sequel, we will refer to this model as Heteroscedastic Latent Variable (HLV) model, to differentiate it from classical latent variable models, where measurement uncertainties features are not explicitly accounted for.

Given the above model structure, parameter estimation is conducted from the probability density function for x under the conditions outlined above, which is a multivariate normal distribution with the following form:

$$x(k) \sim N_n\left(\mu_X, \Sigma_x(k)\right) \tag{7.3}$$

with

$$\begin{aligned} \Sigma_x(k) &= \Sigma_l + \Delta_m(k) \\ \Sigma_l &= A\,\Delta_l\,A^T \end{aligned} \tag{7.4}$$

The likelihood function for a reference data set, composed by n_{obs} multivariate observations, is then given by:

$$\begin{aligned} L(\mu_X, \Sigma_l) &= \prod_{k=1}^{n_{obs}} \left\{ \frac{1}{(2\pi)^{n/2}\left|\Sigma_x(k)\right|^{1/2}} \exp\left[-\frac{1}{2}(x(k)-\mu_X)^T\,\Sigma_x^{-1}(k)\,(x(k)-\mu_X) \right] \right\} \\ \Sigma_x(k) &= \Sigma_l + \Delta_m(k) \end{aligned} \tag{7.5}$$

Therefore, the log-likelihood function, in terms of which calculations are actually conducted, is (C stands for a constant):

$$\begin{aligned} \Lambda(\mu_X, \Sigma_l) &= -\frac{n \cdot n_{obs}}{2}\ln(2\pi) - \frac{1}{2}\sum_{k=1}^{n_{obs}} \ln\left|\Sigma_x(k)\right| - \frac{1}{2}\sum_{k=1}^{n_{obs}} \left[(x(k)-\mu_X)^T\,\Sigma_x^{-1}(k)\,(x(k)-\mu_X) \right] \\ &= C - \frac{1}{2}\sum_{k=1}^{n_{obs}} \ln\left|\Sigma_x(k)\right| - \frac{1}{2}\sum_{k=1}^{n_{obs}} \left[(x(k)-\mu_X)^T\,\Sigma_x^{-1}(k)\,(x(k)-\mu_X) \right] \end{aligned} \tag{7.6}$$

Parameter estimates are then found from those elements of the parameter vector $\theta = \left[\mu_X^T, vec\left(\Sigma_l\right)^T \right]^T$ that maximize the log-likelihood function:

$$\hat{\theta}_{ML} = \max_{\theta} \Lambda\left(\theta \middle| \left\{ x(k), \sigma_m(k) \right\}_{k=1,n_{obs}} \right) \tag{7.7}$$

In fact, the situation is more involved, as Σ_l has certain *a priori* properties that should be satisfied also by its estimate, $\hat{\Sigma}_l$, namely that it should be both symmetric and non-negative definite (Rao, 1973). During the course of our work, several approaches to solve (7.7) were tried out, with different degrees of enforcement of the restrictions arising from symmetry and non-negative definiteness. The one that provided more consistent performance is based upon the (usual) assumption that latent variables have a diagonal covariance matrix, Δ_l, the coefficient matrix A being estimated according to a procedure similar to the one adopted in Wentzell *et al.* (1997a). In this procedure, we start from an initial estimate, A_0, and the numerical optimization algorithm proceeds by finding the optimal rotation matrix R, defined by angles $\underline{\alpha} = \left[\alpha_1 \; \alpha_2 \cdots \alpha_{n-1}\right]^T$, that maximizes (along with the reminding parameters, $\hat{\Delta}_l$ and $\hat{\underline{\mu}}_X$) objective function (7.7):

$$\hat{A} = R(\underline{\alpha})\hat{A}_0 \tag{7.8}$$

$$R(\underline{\alpha}) = R_1(\alpha_1) \cdot R_2(\alpha_2) \cdot \cdots \cdot R_{n-1}(\alpha_{n-1}) \tag{7.9}$$

where,

$$R_1(\alpha_1) = \begin{bmatrix} \cos\alpha_1 & -\sin\alpha_1 & 0 & \cdots & 0 \\ \sin\alpha_1 & \cos\alpha_1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \; R_2(\alpha_2) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & \cos\alpha_2 & -\sin\alpha_2 & \cdots & 0 \\ 0 & \sin\alpha_2 & \cos\alpha_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \; \textit{etc.} \tag{7.10}$$

As $\hat{\Sigma}_l = \hat{A}\hat{\Delta}_l\hat{A}^T$ (from the invariance property of the maximum likelihood estimators; Montgomery & Runger, 1999), the symmetry property is automatically satisfied. The

non-negative definiteness property is enforced by indirect calculation of $\hat{\Delta}_l$, through[29] $\hat{\Delta}_l = \hat{\Delta}_l^{*1/2} \cdot \hat{\Delta}_l^{*1/2}$. Under these considerations, the optimization problem to be solved remains an unconstrained one, and we have used a gradient optimization algorithm to address it. Gradients are given by the following set of equations, for which the complete deduction can be found in Appendix B (see also this Appendix for nomenclature details):

$$
\begin{aligned}
D_{\underline{\mu}_X}\Lambda(\underline{\mu}_X,\underline{\lambda},\underline{\alpha}) &= \frac{1}{2}\sum_{k=1}^{n_{obs}}\left\{\Delta x(k)^T\left(\Sigma_x(k)^{-1}+\left(\Sigma_x(k)^{-1}\right)^T\right)\right\} \\
D_{\underline{\lambda}}\Lambda(\underline{\mu}_X,\underline{\lambda},\underline{\alpha}) &= -\frac{1}{2}\sum_{k=1}^{n_{obs}}\left\{2\left\{vec\left(\Sigma_x(k)^{-1}\right)^T\right\}^T N_n(A^*\otimes I_n)\left(I_p\otimes A\right)\mathbf{T}\right\}+ \\
&\quad +\frac{1}{2}\sum_{k=1}^{n_{obs}}\left\{2\Delta x(k)^T\otimes\Delta x(k)\left[\left(\Sigma_x(k)^T\right)^{-1}\otimes\Sigma_x(k)^{-1}\right]N_n(A^*\otimes I_n)\left(I_p\otimes A\right)\mathbf{T}\right\} \\
D_{\underline{\alpha}}\Lambda(\underline{\mu}_X,\underline{\lambda},\underline{\alpha}) &= -\frac{1}{2}\sum_{k=1}^{n_{obs}}\left\{2\left\{vec\left(\Sigma_x(k)^{-1}\right)^T\right\}^T N_n(A^*\otimes I_n)\left[\left(\Delta^{1/2*}\right)^T\otimes I_n\right]\left(A_0{}^T\otimes I_n\right)\tilde{\mathbf{G}}\right\}+ \\
&\quad +\frac{1}{2}\sum_{k=1}^{n_{obs}}\left\{2\Delta x(k)^T\otimes\Delta x(k)\left[\left(\Sigma_x(k)^T\right)^{-1}\otimes\Sigma_x(k)^{-1}\right]N_n(A^*\otimes I_n)\left[\left(\Delta^{1/2*}\right)^T\otimes I_n\right]\left(A_0{}^T\otimes I_n\right)\tilde{\mathbf{G}}\right\}
\end{aligned}
\tag{7.11}
$$

7.2 Relationship with other Latent Variable Models

Model (7.1)–(7.2) presents some structural similarities with other latent variable formulations that deserve closer analysis. In particular, we will look more carefully to the statistical models underlying MLPCA (Wentzell *et al.*, 1997a) and the classical factor analysis model, FA (Jobson, 1992; Johnson & Wichern, 1992, 2002).

[29] The * is to emphasize that $\hat{\Delta}_l^{*1/2}$ is not the usual square root of a symmetrical positive definite matrix (see Johnson & Wichern, 1992, p. 53), as it does not necessarily need to be positive definite.

PCA models

Principal components analysis (PCA, see also Appendix D) is commonly used in practice as a technique for dimensionality reduction in exploratory data analysis (EDA), or as a "pre-processing" step in regression analysis (as in PCR), to handle the collinearity problem (Jackson, 1991). Algorithmically, it is the solution of an optimization problem, that consists of finding out the set of mutually orthogonal linear combinations of original variables (with coefficients constrained to unit norm), that maximize the residual variability left in data after the portion explained by the former linear combinations has been removed. This task only requires knowledge of the first and second order moments for the set of variables (which are usually assumedly multivariate *iid*), i.e., their mean and covariance matrix (in practice an estimate of them is used), and it can be proven that the optimal linear combinations (loading vectors) are the normalized eigenvectors of the covariance matrix. Thus, PCA does not necessarily have to be considered a model as such in the usual sense, but often as a multivariate statistical analysis technique, with the above mentioned properties. Assuming no principal components are disregarded, then the covariance matrix can be written in terms of the loading vectors and their associated eigenvalues, in the following way:

$$\Sigma = P\Delta P^T \tag{7.12}$$

where Σ is the covariance matrix, P the matrix with the loading vectors in its columns, and Δ a diagonal matrix with the eigenvalues associated with loading vectors along the diagonal. From (7.12) we can see that the "full-dimensional" PCA corresponds to the well known spectral decomposition of a symmetric matrix, in this case the covariance matrix.

From a different perspective, Johnson & Wichern (1992) present another optimal result for PCA, regarding its approximation capability. According to this result, the PCA subspace (i.e., the space spanned by the above referred linear combinations) is the one where the projections of all data points minimize the sum of squares of residuals (given by the difference of the original data and their projections). Wentzell *et al*. (1997a) follow a similar approach when developing MLPCA (see also Section 5.1.3), by using a statistical description for the measured data matrix, say X, based on a PCA-type inner

structure to be estimated when all the measured variables are subject to Gaussian errors with known uncertainties. Such a description relies on the assumption that the matrix of measured data, X ($n_{obs}\times n$), arises directly from the values of the underlying latent variables, L ($n_{obs}\times p$), through a coefficient matrix, A ($n\times p$), to which measurement noise is added (the entries of E, $n_{obs}\times n$):

$$X = LA^T + E \tag{7.13}$$

The fundamental assumptions made by Wentzell *et al.* (1997a), regarding the above model, are that: (i) there exists a true underlying p-dimensional model, given by (7.13), for a data matrix, (ii) deviations from this model come only from measurement random errors and (iii) these random errors are normally distributed around the true measurement values, with general, *but known*, standard deviations and covariance. When model (7.13) is estimated using a maximum likelihood approach, it redounds in the conventional PCA decomposition of X when measurements are uncorrelated with equal variances, but provides better estimates when errors have more complex error structures (Wentzell *et al.*, 1997b). Considering the particular case where measurement errors do not exhibit any correlation structure along rows or columns of E, we can write down the model for each row (i.e., for each multivariate observation) and compare it to (7.1)–(7.2):

$$\begin{aligned} x(k) &= A\cdot l(k) + \underline{\varepsilon}_m(k) \\ \underline{\varepsilon}_{\mathrm{m}}(k) &\sim iid\ MN_n\left(0, \Delta_m(k)\right) \end{aligned} \tag{7.14}$$

Comparing (7.14) with (7.1)–(7.2), we can see an apparent structural similarity, the most fundamental difference being that (7.14) does not model any underlying statistical behaviour for the latent variables, as happens in (7.1)–(7.2). Similarly to Wentzell *et al.* (1997a), a maximum likelihood type of approach for estimating the relevant parameters is adopted here, but the likelihood function is necessarily different. The higher number of parameters to estimate and the objective function non-linearity for the proposed

formulation will naturally have consequences regarding computation time and the dimensionality of the problems that one can cope with.

Another situation where a PCA model often arises is in the scope of MSPC based on a latent variable framework (LV-MSPC). Traditionally, such MSPC procedure is implemented through a T^2 statistic (that measures variation within the PCA model) and the Q or *SPE* statistic (that measures the amount of variation *not captured* by the PCA model) (Jackson & Mudholkar, 1979; MacGregor & Kourti, 1995). What is important to point out here, and is frequently overlooked, is the statistical model underlying such a procedure. The statistics used in this methodology were derived assuming the set of random variables involved $x = [x_1, x_2, \cdots x_n]^T$ to follow a multivariate normal process with mean $\mathbf{0}$ and *full rank* covariance matrix Σ (Jackson & Mudholkar, 1979).[30] This assumption of full rank n for the covariance matrix is used in the derivation of the approximate distribution for the Q statistic (although the result still applies when some of the latent roots are zero, if we replace the summations up to n by summations until the maximum number of non-zero latent roots). When performing LV-MSPC based on PCA, only an adequate number of components is retained, which explain the normal "structural" variation, say p, and it is *considered* that, for all practical purposes, the variability explained by the reminding factors can be neglected, corresponding to "unstructured" variation (Chiang *et al.*, 2001). Therefore, this approach corresponds to the following statistical model:

$$\begin{aligned} x(k) &= \underline{\mu}_X + Al(k) + A_R l_R(k) \\ &= \underline{\mu}_X + Al(k) + R(k) \\ l(k) &\sim iid\ MN_p(\underline{0}, \Delta_S) \\ R(k) &\sim iid\ \ Mpdf_{n-p}(\cdot) \end{aligned} \tag{7.15}$$

[30] In fact, this strictly applies to the Q statistic, because the well known T^2 statistic can be used in quite general contexts, e.g. in MSPC *without* any latent variable formalism, but also, in particular, in the present situation.

Thus, as referred by Jackson (1991), the variation of x is the result of the cumulative contributions coming from: (i) the multivariate mean (first term on the right side of the equation); (ii) structured variation (second term); and (iii) unstructured or unexplained variability (third term). Comparing (7.15) with (7.1)–(7.2), it is possible to see again a remarkable structural similarity between the models, that is also extensive to some degree to the distributional assumptions. This is not surprising, as they appear in the same context, but posing different assumptions regarding the unstructured component. In particular, in (7.15) a given structure is being implicitly "enforced" to the unexplained variability, $R(k)$, namely that it should be orthogonal to the explained variability. This seems to be a reasonable assumption, but still brings as a natural consequence the conditioning of the nature of the multivariate distribution underlying $R(k)$ (here generally referred to as a multivariate probability density function, *Mpdf*).

On the other hand, in (7.1)–(7.2), although we assume a given statistical distribution for the measurement uncertainty part, this is by no means related to the description adopted for the structured part. However, it is also possible to add to our model an additional term, say $\underline{\varepsilon}_u$, that accounts for the portion of variability regarding modelling mismatch and unaccounted disturbances, statistically described in similar terms to $R(k)$.

Factor analysis (FA) models

Factor analysis (FA) is a multivariate technique with some connections to PCA, but also with some fundamental differences (Jackson, 1991). FA was designed to explain the *cross-correlation* structure between all variables, assuming a well specified statistical model. This means that the explained structure only holds to the extent that the validity of the assumed model is verified. Thus, there are already two important differences worth emphasizing: PCA aims to effectively explain *variability* (not *correlation*), and does not necessarily have to rely on any underlying statistical model. However, it is useful to adopt the basic PCA model structure used for conducting LV-MSPC, equation (7.16), in order to better understand the assumptions made in FA:

$$x(k) = \underline{\mu}_X + Al(k) + R(k) \tag{7.16}$$

As both $l(k)$ and $R(k)$ are random variables, the covariance of x is given by

$$\begin{aligned}\Sigma_x &= A\Delta_l A^T + E \\ &= A^* A^{*T} + E\end{aligned} \tag{7.17}$$

with $A^* = A\Delta_l^{1/2}$ ($\Delta_l^{1/2}$ is a diagonal matrix with the square root of the elements of Δ_l). Now, the FA model begins with a similar structure:

$$x(k) = \underline{\mu}_X + Al(k) + \varepsilon(k) \tag{7.18}$$

but inserts additional constraints in the expression for the covariance of x, through the specification of the following statistical model for all the random variables:

$$\begin{aligned}&E\left(l(k)\right) = \underline{0} \\ &\text{cov}\left(l(k)\right) = I, \text{ the identity matrix} \\ &E\left(\varepsilon(k)\right) = \underline{0} \\ &\text{cov}\left(\varepsilon(k)\right) = \Psi, \text{ a diagonal matrix} \\ &l(k) \text{ and } \varepsilon(k) \text{ are independent}\end{aligned} \tag{7.19}$$

Therefore, the covariance matrix underlying FA models has the following form:

$$\Sigma_x = AA^T + \Psi \tag{7.20}$$

The most evident difference between (7.17) and (7.20) regards the covariance term arising from the statistical behaviour reserved for the residuals, which for FA is now a diagonal matrix, representing the unique contributions to the overall covariance arising from each variable. Thus, once the relevant latent variables (factors in the FA nomenclature) are selected, no residual covariance structure should remain. This leads

us to another difference between PCA and FA: while in PCA we keep adding components until residual variances are sufficiently low, in FA we do such until the residual covariance has been sufficiently reduced.

Expression (7.20) plays a central role in FA, as it defines the assumed structure for the covariance matrix of x, whose explanation is the main goal to be achieved by this technique. FA proceeds by estimating the parameters involved in A and Ψ, and there are several methods for doing so (Jackson, 1991; Jobson, 1992; Johnson & Wichern, 1992, 2002). However, the model is still under-defined, as there are still multiple possible solutions for (7.20). Thus, in order to eliminate this inherent indeterminacy, one has to provide additional constraints to remove degrees of freedom.

Comparing the FA model to the proposed one, (7.1)–(7.2), and in particular its expression for the covariance of x, (7.4), it is possible to verify that the proposed approach for estimating the parameters leads to an analogous "common factor" term, and in this sense is quite similar to FA, but the residual or unstructured part of the model resembles more the one adopted in PCA, although extended to incorporate heteroscedasticity. Thus, we might say that our proposed model lies somewhere between FA and PCA, with heteroscedastic formulations (hence the designation of our model as *heteroscedastic latent variable* model, HLV). Furthermore, we can say that, with some minor modifications to the methodology, a maximum likelihood heteroscedastic FA model can also be put forward through the inclusion of an additional term, $\underline{\varepsilon}_u(k)$, regarding unique contributions from each variable, $x(k) = \underline{\mu}_X + A \cdot l(k) + \underline{\varepsilon}_m(k) + \underline{\varepsilon}_u(k)$, which would imply adding a diagonal covariance matrix to the expression for the covariance of x, that would then become $\Sigma_x(k) = A \Delta_l A^T + \Delta_m(k) + \Psi$.

7.3 HLV – MSPC Statistics

In this section we present the monitoring statistics and discuss some issues regarding the implementation of MSPC within the scope of the HLV model, formulated and discussed in the previous sections. Efforts were directed towards developing statistics that would be analogous to their well known counterparts, i.e., to T^2 and Q for MSPC based on PCA (Wise & Gallagher, 1996).

7.3.1 Monitoring Statistics

Conventional T^2 and Q statistics were designed to follow the behaviour of the two random components present in a PCA model: one reflecting the structured variation arising from latent variables sources, which is "followed" by the T^2 statistic, and the other relative to the unstructured part, driven by the residuals, followed by the Q statistic. Since our proposed model also contains structured and unstructured components, the same rational will be pursued. The structured, or "within" latent variables subspace variability, will be monitored in the original variable domain, instead of the latent variable domain (as done in PCA-MSPC), in order to account for the effects of the (known) measurement uncertainties. This leads to the definition of the following statistic:

$$\begin{aligned} &T_w^{\,2}(k) = \left(x(k) - \mu_X\right)^T \Sigma_x^{\,-1}(k)\left(x(k) - \mu_X\right) \\ &\Sigma_x(k) = \Sigma_l + \Delta_m(k) \\ &\left(\Sigma_l = A\,\Delta_l\,A^T\right) \end{aligned} \tag{7.21}$$

where $x(k)$ represents the k^{th} measured multivariate observation, and the other quantities keep the same meaning as before. It follows a $\chi^2(n)$ distribution, n being the number of variables. $T_w^{\,2}(k)$ considers simultaneously the variability arising from both the structured (process) and unstructured (measurement noise) variability. Let us now define the statistic Q_w, that considers only the unstructured part of the HLV model, say $r(k)$, associated with measurement noise:

$$\begin{aligned} &Q_w = r^T(k)\Delta_m^{\,-1}(k)r(k) \\ &r(k) = x(k) - \mu_X - Al(k) = \underline{\varepsilon}_m(k) \end{aligned} \tag{7.22}$$

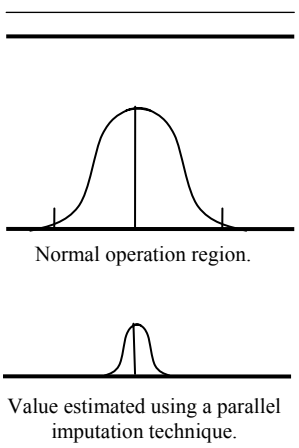
which follows a $\chi^2(n-p)$ distribution, with n and p being the number of variables and latent variables (pseudo-rank), respectively. In practice, the true values for the above quantities are unknown, and those that maximize the log-likelihood function will be

used as their estimates. Furthermore, $l(k)$ values are calculated using non-orthogonal (maximum likelihood) projections (Wentzell *et al.*, 1997b), given by:

$$\hat{l}_{ML}(k) = \left(\hat{A}_{ML}{}^{T} \Delta_m{}^{-1}(k) \hat{A}_{ML} \right)^{-1} \hat{A}_{ML}{}^{T} \Delta_m{}^{-1}(k) \left(x(k) - \hat{\mu}_{X,ML} \right) \tag{7.23}$$

7.3.2 Missing Data

The incorporation of uncertainty information regarding each measured value in HLV-MSPC not only adds a new important dimension to it, but also brings some parallel advantages. One of them is the inherent ability to handle reasonable amounts of missing data, in a coherent and integrated way. Usually, missing data are replaced by conditional estimates obtained under a set of more or less reasonable assumptions, or through iterative procedures where, in practical terms, missing values play the role of additional parameters to be estimated. In the proposed procedure, when a datum is missing, we simply have to assign a value to it, together with its associated uncertainty. This assigned datum can be simply the mean of the normal operation data, with the corresponding standard deviation as an adequate uncertainty value. Alternatively, we can also assign the mean value together with a very large score for its associated measurement uncertainty, the rational being that a missing value is virtually given by any value with an "infinite uncertainty". More precise estimates, obtained through data imputation techniques, can also be adopted if they are able to provide us also with the associated uncertainties (Figure 7.1).



a) No external knowledge: *use any value + "Infinite" uncertainty*

b) Knowledge of normal operation: *use mean value + uncertainty encompassing normal operation region*

c) Knowledge about an estimate of the missing value: *use estimate + estimate uncertainty*

Figure 7.1. Three levels of knowledge incorporation with regard to missing data estimation: a) no external knowledge; b) knowledge about the mean and standard deviation under normal operation conditions; c) imputation of missing data values using a parallel imputation technique.

Alternative procedures for implementing HLV-MSPC, that relax some of the parametric assumptions postulated in the proposed model, are referred in Appendix C.

7.4 Illustrative Applications of HLV-MSPC

In this section the main results obtained from the application of HLV-MSPC to a number of different simulated scenarios, where measurement uncertainties were allowed to vary (heteroscedastic noise), are presented. The case study initiated in Section 4.4.4 will also be concluded here, with the main goal of extracting from it knowledge regarding process variability trends.

7.4.1 Application Examples

The first four examples are based on data generated by the following latent variable model:

$$
\begin{aligned}
x(k) &= 5 \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} \cdot l(k) + \varepsilon_m(k) \\
l(k) &\sim iid\ N(0, \Sigma_l),\ \Sigma_l = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \\
\varepsilon_m(k) &\sim id\ N(0, \Delta_m(k))
\end{aligned}
\tag{7.24}
$$

The measurement noise covariance can vary along time in various ways, as explained below, and each example covers a different scenario regarding time variation of measurement uncertainty. For comparative purposes, the results obtained using classic PCA are also presented. The statistics for PCA-MSPC are denoted by T^2 and Q, and those for HLV-MSPC as T_w^2 and Q_w. All simulations carried out for the different scenarios share a common structure: first, in the training phase, 1024 multivariate observations are generated using model (7.24) in order to estimate the reference PCA and HLV models; then, in the testing phase, 1000 observations of new data are generated, half of which are relative to normal operation (from observations 1 to 500), while the other half corresponds to an abnormal operation situation (observations 501 to 1000). For each of these two parts we calculate MSPC statistics, and the percentage of significant events identified (events above statistical limits), for the significance level adopted ($\alpha = 0.01$). In order to enable for a more sound assessment of results, the testing phase was repeated 100 times, and the performance medians over such repetitions computed. Furthermore, two abnormal situations (faults) are explored in each scenario, as follows:

F1) A step change of magnitude 10 is introduced in all variables;

F2) A structural change in the model is simulated, by modifying one of the entries in the coefficient matrix

$$
\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -0.5 \\ 1 & -1 \end{bmatrix}
\tag{7.25}
$$

Example 1: Constant Uncertainty for the Reference Data (at Minimum Level)

In this example, measurement noise standard deviations for the reference data set (used to define control limits) were kept constant and at the minimum values that will be used during the test phase. For the test data, measurement uncertainties are allowed to vary randomly, according to the uniform distribution $\sigma_m^{X_i}(k) \sim U(2,6)$ (we will refer to this situation as "complete heteroscedasticity"). The corresponding results are presented in Table 7.1, for the two types of faults mentioned above (*F1* and *F2*).

Table 7.1. Median of the percentages of significant events identified in 100 simulations for Example 1, under normal and abnormal operation conditions (*Faults F1 and F2*).

Fault	*Statistic*	*Normal Operation*	*Abnormal operation*
F1	T^2	2.40	17.80
	Q	31.40	79.70
	T_w^2	1.20	27.80
	Q_w	1.00	25.20
F2	T^2	2.30	1.40
	Q	31.60	45.20
	T_w^2	1.20	4.80
	Q_w	1.00	6.80

The PCA's Q statistic detects a very large number of false alarms, whereas T^2 detects almost twice the expected rate under the adopted statistical significance level (0.01). The apparently good performance of Q under abnormal conditions is a consequence of the low statistical limits established, which are related with the low noise reference data used. This leads to a sensitive detection of any fault, but at the expense of a very large rate of false alarms under normal operation. HLV-MSPC statistics perform consistently better, particularly when we compare T_w^2 and T^2 performances.

Example 2: Constant Uncertainty for the Reference Data (at Maximum Level)

Looking now to what happens if uncertainties in the reference data are held constant at the maximum levels used in the test data set (Table 7.2), we can see that the opposite

detection pattern occurs with the T^2 and Q statistics, as expected. In these examples, as the reference data consists of highly noisy measurements, and therefore the control limits are set at higher values, the detection ability for false alarms becomes smaller when noise characteristics change. This also drastically reduces the capability for detecting significant events. Under this situation, HLV-MSPC statistics also outperform their classical counterparts.

Table 7.2. Median of the percentages of significant events identified in 100 simulations for Example 2, under normal and abnormal operation conditions (*Faults F1 and F2*).

Fault	*Statistic*	*Normal Operation*	*Abnormal operation*
F1	T^2	0.40	5.20
	Q	0.00	1.00
	T_w^2	1.00	23.80
	Q_w	1.00	24.00
F2	T^2	0.40	0.20
	Q	0.00	0.00
	T_w^2	0.80	3.80
	Q_w	1.00	6.00

In the previous results, measurement uncertainties for each value of each variable in the test set were allowed to change randomly from observation to observation, according to the probability distribution referred. Scenarios were also tested where the values for all variables in the same row were assumed to have the same uncertainty, and we found out that the same conclusions hold for this situation. For illustrative purposes, Table 7.3 presents the results obtained for fault F1, when the reference data was generated at maximum uncertainty values.

Table 7.3. Median of the percentages of significant events identified in 100 simulations (when the uncertainties for all observations in the same row share the same variation pattern), under normal and abnormal operation conditions (*Fault F1*).

Fault	*Statistic*	*Normal Operation*	*Abnormal operation*
F1	T^2	0.40	4.80
	Q	0.20	1.60
	T_w^2	1.00	28.00
	Q_w	1.10	30.20

Example 3: Variable Data Uncertainty for Reference and Test Sets

The examples mentioned so far address situations where the training set variables have constant measurement uncertainty, whereas the test set uncertainties have heteroscedastic behavior. This mismatch between training and testing situations has serious consequences in the performance of PCA-based MSPC. The following examples explore situations where both the reference and test data were generated under similar conditions of measurement uncertainty heteroscedasticity. First, let us consider the already described situation of complete heteroscedasticity. From Table 7.4, it is possible to see that HLV-MSPC statistics still seem to present the best performance, although PCA-based MSPC counterparts also achieve good scores for normal operation.

Table 7.4. Median of the percentages of significant events identified in 100 simulations (Example 3), under normal and abnormal operation conditions (*Faults F1 and F2*).

Fault	*Statistic*	*Normal Operation*	*Abnormal operation*
F1	T^2	1.00	8.80
	Q	1.40	15.40
	T_w^2	1.00	25.20
	Q_w	1.00	25.00
F2	T^2	1.00	0.80
	Q	1.40	3.40
	T_w^2	1.00	4.60
	Q_w	1.00	6.60

Once again, the above conclusions do not change in the situation where uncertainty for all of the variables does change together, as shown for fault F1 in Table 7.5.

Table 7.5. Results for fault F1, with variable uncertainty both in the reference and test data (when the uncertainties for all observations in the same row share the same variation pattern).

Fault	*Statistic*	*Normal Operation*	*Abnormal operation*
F1	T^2	0.80	9.90
	Q	1.90	13.40
	T_w^2	1.00	28.50
	Q_w	1.00	29.20

Example 4: Handling the Presence of Missing Data

This example explores the capability of the proposed methodology for handling missing data randomly scattered through data sets. The underlying model used to generate noiseless data sets is the same as before (*Example 1*), but some data records were now removed through an automatic random procedure that approximately eliminates a pre-specified percentage of values (it removes *on average* the chosen percentage), here fixed at 10%. As for the previous examples, results presented below regard testing data performances. For HLV-MSPC, two different simple procedures for replacement of missing data were followed:

i. in the first one (*MD I*), the mean for each variable was inserted in a missing datum position, and a high value associated to it at the corresponding position in the uncertainty table (e^{10});

ii. in the second procedure (*MD II*), this estimate was refined, using the available reference data to estimate the mean and standard deviations for each variable, the former being used to replace missing data and the latter one to specify the associated uncertainty.

For PCA-MSPC, missing data estimates were based upon reference data means (*MD*). Table 7.6 presents the results obtained for fault *F1*, with the values for HLV-MSPC and PCA-MSPC for the original data (i.e., without missing data) also being reported. It is possible to verify that there is a sensible and expected decrease of detection performances for HLV-MSPC statistics under the more pessimistic imputation method, *MD I*, which are improved by using procedure *MD II*. From these results we can say that it is still advisable to continue with the implementation of HLV-MSPC in the presence of missing data, as the results with missing data are in general superior to those of PCA-MSPC *without* missing data.

Table 7.6. Median of the percentages of significant events identified in 100 simulations (Example 4), under normal and abnormal operation conditions (fault F1).

Statistic	*Operation*	PCA (orig)	PCA (MD)	HLV (orig)	HLV (MD I)	HLV (MD II)
T^2	*Normal*	1.10	0.80	1.00	0.80	1.20
	Abnormal (F1)	10.80	8.40	31.90	25.80	27.80
Q	*Normal*	2.80	5.70	1.20	0.80	1.20
	Abnormal (F1)	18.80	24.10	32.20	24.60	28.00

7.4.2 Analysis of Pulp Quality Data

In this section we apply our HLV-MSPC procedure to the pulp and paper quality data set projected at scale $j = 3$ (corresponding to "averages" over $2^3 = 8$ days), in order to take full advantage of the uncertainty information that the proposed MRD framework puts at our disposal. These uncertainty profiles, for the approximation coefficients regarding the nine variables studied along the time index (at scale $j = 3$), are represented in Figure 7.2. Since all of these variables are derived from the plant quality control laboratory, their acquisition periodicity is almost the same, and therefore their profiles do exhibit similar patterns.

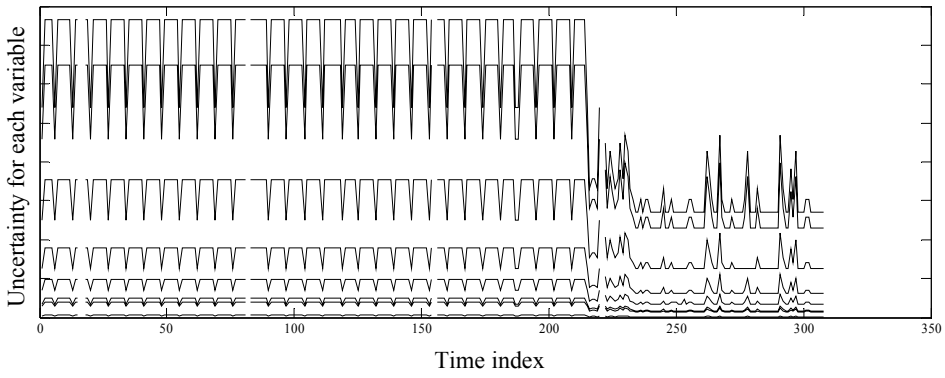


Figure 7.2. Patterns of data uncertainty variation along time index for the 9 pulp quality variables analyzed (data is aggregated in periods of 8 days, and such time periods are reflected by the time index shown here).

A Phase I study was conducted, and the HLV-MSPC statistics computed in order to analyze the variability structure across time. For setting the pseudo-rank parameter, a first guess can be easily provided by applying classical PCA to our data and then using

one of the associated selection procedures available for identification of the proper number of PC to retain (Dable & Booksh, 2001; Meloun *et al.*, 2000; Qin & Dunia, 2000; Thomas, 2003; Valle *et al.*, 1999; Vogt & Mizaikoff, 2003; Wold, 1978). This initial guess can then be tested and revised in pilot implementations of the method over real data. A final selection should also be validated against the values of the diagonal matrix, Δ_l, estimated from such implementations, in order to check if they are also consistent with such a choice. In the present case study, this parameter was set as $p = 3$. Figure 7.3 illustrates the values obtained for the T_w^2 statistic, where it is possible to identify a process shift after time instant 240, occasionally spiked with some rare but very significant abnormal events. For comparison purposes, we also present, in Figure 7.4, the values obtained for the analogous T^2 statistic, obtained by conducting the same analysis using PCA-MSPC, where the sustained shift in the last period of time almost passes undetected, whereas high data variability present in the beginning (where uncertainties have higher values) is not properly down-weighted, leading to an inflated variation pattern.

The T_w^2 profile provides a rough vision over the conjoint time behaviour, but it is possible to zoom into it (without having to analyze the variables separately, in which case we would be missing any changes in their correlation structure), by looking to what happens to the HLV scores provided by equation (7.23), as shown in Figure 7.5. From these plots, it is possible to identify several trends affecting the three scores: a long range oscillatory pattern for the first score, a decreasing trend with shorter cyclic patterns superimposed for the second score, and a stable pattern that begins to oscillate in the final periods of time for the third score.

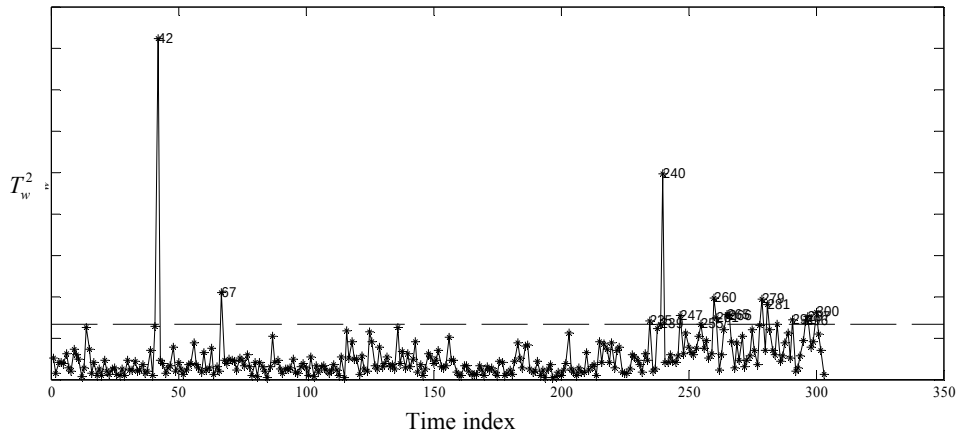


Figure 7.3. HLV-MSPC: values for the T_w^2 statistic in the pulp quality data set.

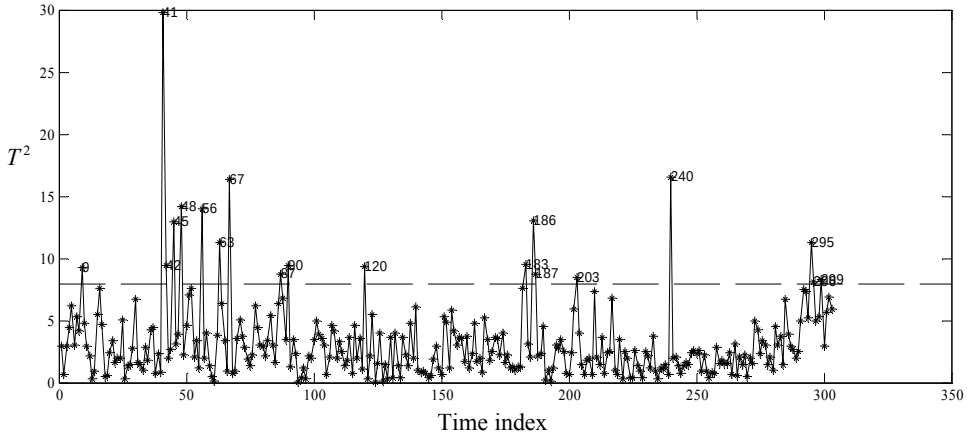


Figure 7.4. PCA-MSPC: values for the T^2 statistic in the pulp quality data set.

By looking into the variables that are responsible for such behaviours, namely through contribution plots for the scores, we can get more insight into the nature of these disturbances, and, eventually, about their root causes. Even though a detailed discussion is beyond the scope of this thesis, one should notice that these types of trends are common in pulp and paper quality data, and can be attributed to issues ranging from seasonal wood variability and harvesting cycles to wood supply policies.

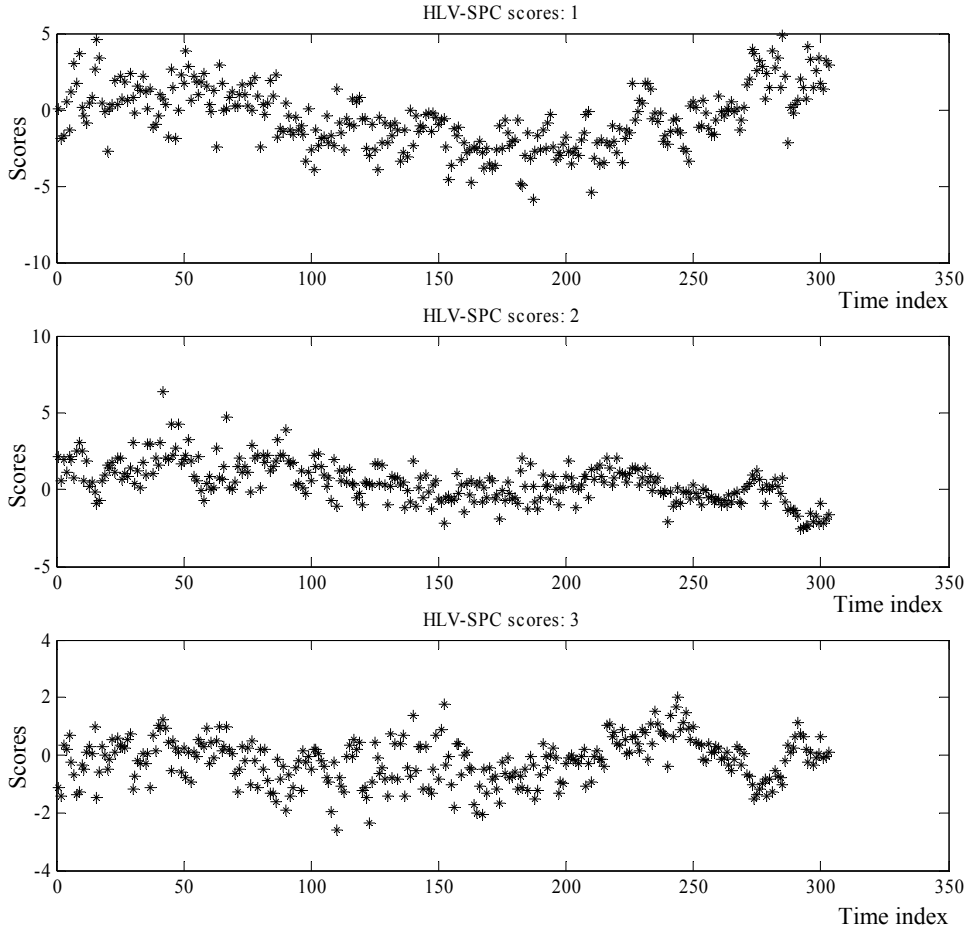


Figure 7.5. HLV scores for the pulp quality data set.

7.5 Discussion

The approach proposed in this chapter was designed to perform SPC under noisy environments, i.e., scenarios where the signal to noise ratio (or, more adequately, signal to *uncertainty* ratio) is rather low, and, furthermore, where the magnitude of the uncertainty affecting each collected value can vary across time. Not only standard measurement systems that conform to the underlying statistical model are covered by this approach (e.g. laboratory tests, measurement devices), but also any general procedure for obtaining data values with an associated uncertainty (e.g. computational calculations, raw material quality specifications, etc.) may be eligible. The added value of our proposed approach increases when the signal variation to uncertainty ratio

becomes smaller. Therefore, it provides an alternative to PCA-MSPC for applications where low signal to noise ratios tend to happen.

The better capability of the proposed approach to estimate the underlying true data subspace was also analyzed through a simulation study. Noiseless data were generated using the model described in Example 1, and then corrupted with noise, whose measurement uncertainties vary randomly between 2 and 6 (uniform distribution). For each trial, 100 multivariate observations were used to estimate the underlying latent variable subspace using classical PCA and HLV. The angle that these estimates make with the true subspace, as well as the respective distances (Golub & Van Loan, 1989) and the Krzanowski similarity factor (Krzanowski, 1979) between the estimated and the true subspaces, were calculated: ANG(PCA), ANG(HLV), DIST(PCA), DIST(HLV), SIMIL(PCA) and SIMIL(HLV), respectively. The Krzanowski similarity factor is a measure of the similarity between two PCA subspaces, ranging from 0 (no similarity) to 1 (exact similarity). The means and standard deviations for these quantities, derived from 100 trials, are presented in Table 7.7, along with the values of the t-statistic for paired t-tests between PCA and HLV results, and the respective p-values. A highly significant better estimation performance in favour of the HLV procedure was thus obtained.

Table 7.7. Mean and standard deviation of the results obtained for the angle, distance and similarity factor between the estimated subspace and the true one, using PCA and HLV (first row). Paired t-test statistics for each measure, regarding 100 simulations carried out, along with the respective p-values (second row).

	ANG(PCA) (º)	ANG(HLV) (º)	DIST(PCA)	DIST(HLV)	SIMIL(PCA)	SIMIL(HLV)
Mean *(Standard dev.)*	26.62 (3.58)	17.23 (2.63)	0.42 (0.06)	0.30 (0.04)	0.91 (0.02)	0.95 (0.01)
t statistic *(p-value)*	29.84 ($<< 10^{-5}$)		30.54 ($<< 10^{-5}$)		-25.89 ($<< 10^{-5}$)	

7.6 Conclusions

In this chapter, an approach for performing SPC in multivariate processes that explicitly incorporates measurement uncertainty information, was presented and discussed. A statistical model was defined and statistics analogous to T^2 and Q derived, that allow

for monitoring both the within model variability as well as the variability around the estimated model. The proposed approach is also able to handle the presence of moderate amounts of missing data in a simple and consistent way.

Preliminary results point out in the direction of advising the use of this framework when measurement uncertainties are available and significant noise affects process measurements. So far, this approach was implemented and tested in examples that do cover dozens of variables. In even larger scale problems the computational load associated with it may become an issue, but we may still apply the same methodology over a subset of variables, where heteroscedasticity is believed to be more critical.

Part IV-B

Multiscale Data Analysis

This is not something imposed by the mathematicians; it came from engineering.

Yves Meyer (1939-), French mathematician (about wavelets).

Chapter 8. Multiscale Monitoring of Profiles

In the last three chapters of Part IV-A, attention has been devoted to several data analysis tasks conducted at a single scale, that take explicitly into consideration the available knowledge regarding noise features, namely data uncertainty. In particular, these approaches are suited for implementation over data previously processed by MRD frameworks (Chapter 4), even though these ones can also be used in more general contexts, namely in multiscale data analysis. We move now to a different application context, aimed at extracting and handling the *multiscale features* present in collected data, in order to come up with improved *process monitoring* procedures.

Two rather different monitoring contexts will be covered. The first one, to be addressed in this chapter, is relative to multiscale monitoring of profiles (Section 2.5.3) whose structural (deterministic) and stochastic properties remain *stationary* (in the length domain) within a sample profile obtained under normal operation conditions. These are the kind of profiles that are acquired from approximately homogeneous production unities, in which stochastic properties, as well as deterministic structures, remain constant, in a general sense, throughout the two dimensional space. They will be named here as *stationary profiles*. The main application scenario envisioned for this class of approaches regards multiscale monitoring applications in the "length domain", and special attention will be given to its application for monitoring of paper surface. As the samples can have different origins and span different portions of surface, the specific spatial location of events in the length domain is not relevant, but rather their

localization in the frequency domain. This is the essence of the *stationary* assumption in the length domain.

In the second context, covered in the following chapter (Chapter 9), our focus will rely in performing multiscale monitoring in the "time domain". Such methodologies deal with patterns often present in industrial data, as a consequence of a wide range of possible process upsets with rather different characteristics of location and localization in the time/frequency domain, as well as due to the intrinsically complex nature of plants, that are usually convoluted networks of processing elements with quite different dynamic characteristics, all of them interfering with variability in quality features of the final manufactured product, but at different scales of time.

In both of these chapters we will assume that no missing data is present and homoscedastic uncertainties. To handle the more complex situations where one or both of these assumptions are not valid, multiscale monitoring approaches can also be developed, for instance, by adequately combining MRD and multivariate data analysis frameworks such as MLPCA, that can handle quite well the type of data structures provided by uncertainty-based MRD frameworks, but this topic will not be covered in the present thesis, being deferred to future work (Chapter 11).

8.1 Description

The general subject of profile monitoring was already introduced in Section 2.5.3, along with a reference to several developments in this field, including some multiscale approaches that have already been proposed. However, the general application scenario that the present methodology is aimed to deal with, presents some particularities, that call for a different monitoring approach. In particular, we are here interested in monitoring profiles that can be sampled at random from a product (already finalized or still under processing), where the *location* of a given feature in the length direction (say X-axis) is not critical, but only the *global* behaviour of the profile obtained for the *relevant scales*. Furthermore, flexibility regarding the incorporation of existent background engineering knowledge should also be allowed, namely in the selection of those scales of interest for each particular phenomenon to be monitored, or in the key profile monitoring features to follow.

The main steps that compose the proposed approach for conducting Multiscale SPC of stationary profiles are summarized in Table 8.1.

Table 8.1. Basic elements of the proposed general methodology for multiscale monitoring of stationary profiles.

1. Acquisition of profile.
2. Multiscale decomposition of the de-trended profile (i.e., the profile with the linear trend removed), obtaining wavelet coefficients at each scale ($j = 1 : J_{dec}$, where J_{dec} is the decomposition depth).
3. Selection of relevant scales for monitoring relevant profile's phenomena.
4. Using only those scales whose indices are relative to phenomena under analysis, calculate the parameters that summarize the relevant information for product quality control purposes (this may require the separate reconstruction of profiles relative to each phenomenon back into the original domain, by applying the inverse wavelet transform to a set of processed coefficients, where the only non-zero elements are those corresponding to the selected scales for each phenomenon).
5. Implementation of SPC procedures for monitoring the parameters calculated in step 3.
6. If an alarm is produced, check its validity and look for root causes when appropriate. Otherwise, return to Step 1, and repeat the whole procedure for the next profile acquired.

Step 3 allows for the incorporation of external knowledge in the process of selecting those scales that are relevant for monitoring, as well as on defining the relevant monitoring statistics.

The overall procedure essentially consists of applying a bank of quadrature mirror filters (Strang & Nguyen, 1997; Vetterli & Kovačević, 1995) that basically works as bandpass filters (Oppenheim *et al.*, 1999), dividing the frequency domain in octave bands, [31] i.e., bands whose ranges increase proportionally to the mean frequency covered, a scheme also known as "constant-*Q*", or constant relative band in the signal processing community (Rioul & Vetterli, 1991).

This organization of the frequency domain turns out to be very adequate for data analysis, since what is usually of interest is the band relative width. For instance, a 10 Hz width band can be considered as a quite narrow one when located around the frequency of 20 000 Hz, but already quite large when centered at 20 Hz. This means that its information content can be much more critical at these lower frequencies than at the higher frequency bands. Therefore, a sound way for organizing frequency information is to pack it into regions of equal relative bandwidth, and not in regions of equal bandwidth, as done through the Short Time Fourier Transform (Vetterli & Kovačević, 1995) or Windowed Fourier Analysis (Mallat, 1998). As a matter of fact, the recognition that such frequency packing could also simplify the analysis of contributions arising from the different parts of the spectrum was already referred in the literature, namely in profilometry applications (Wågberg & Johansson, 2002), without explicitly addressing wavelet transforms.

Due to the *stationary* assumption, the location property of the wavelet transform in the length domain is not relevant, and therefore the analysis carried out is "global" in this domain, but "local" in the frequency domain, where phenomena occurring at different scales are analyzed and monitored separately.

In the next section the case study where this methodology is applied is introduced. It concerns monitoring of the paper surface, and the measurement technique adopted to provide raw measurements of the paper surface, profilometry, is also presented. The

[31] This term is borrowed from the musical nomenclature, meaning an entire sequence of eight notes (therefore the term "octave") during which the frequency doubles when going from the first note to the last one, i.e., frequency doubles each time we go up an "octave". Furthermore, not only the frequency doubles, but the same happens with the ranges of frequencies covered by each octave.

measurement device used has a built-in functionality of providing summary statistics for both surface waviness and roughness phenomena (two types of surface irregularities occurring at different length scales or frequency bands), which is also explored for predicting paper quality from the stand point of the final user. However, the proposed monitoring approach is entirely based on the raw profiles, in order to take the most out of our multiscale analysis framework. Results regarding its application to simulated scenarios, as well as real industrial profiles, are presented in the following section. Final conclusions are drawn in the third section.

8.2 Case Study: Multiscale Monitoring of Paper Surface

8.2.1 Paper Surface Basics

Paper is a very complex material, exhibiting properties that derive from a structural hierarchy of arrangements for different elements (molecules, fibrils, fibres, network of fibres, etc.), beginning at a scale of a few nanometres and proceeding all the way up to a few dozens centimetres or even meters (Table 8.1; Kortschot, 1997).

Table 8.2. The multiscale structure of paper (based on Kortschot, 1997).

Scale	***Structural Component***
1 *nm* – 10 *nm*	Molecular structure and packing: • *Cellulose* • *Hemicellulose* • *Lignin* • *Other*
10 *nm* – 1 μm	Internal structure of the fibre: • *Softwood tracheids* • *Hardwood fibres* • *Hardwood vessels* • *Ray cells* • *Compression wood* • *Tension wood*
1 μm – 10 *mm*	Fibre morphology: • *For different types of fibres (softwood tracheids, hardwood fibres, hardwood vessels, ray cells, fines)*
1 μm – 10 *mm*	Paper microstructure
1 *mm* – 10 *cm*	Paper mesostructure
5 *mm* – 30 *cm*	Paper macrostructure

This complexity is also present at its boundary, the paper surface, which plays a central role in many of the relevant properties from the perspective of the end user, such as general appearance (optical properties, flatness, etc.), printability (e.g. absorption of ink) and friction features, to name a few (Kajanto *et al.*, 1998). Being aware of this importance, the Pulp and Paper Industry developed methods to assess and characterize paper surface features at different scales, and, in particular, special attention has been devoted to surface phenomena known as *roughness* and *waviness*.

Roughness is a fine length-scale phenomenon, that results from the superposition of the so called optical roughness (scales up to $1\,\mu m$), micro-roughness (scales between $1\,\mu m - 100\,\mu m$) and macro-roughness (scales between $0.1\,mm - 1\,mm$), each one with their own specific structural elements (Kajanto *et al.*, 1998):

- *Optical roughness* is related with individual pigment particles and pulp fibres;
- *Micro roughness* is mainly concerned with the shapes and positioning of fibres and fines in the network structure;
- *Macro roughness* is related to paper formation.[32]

Roughness is usually characterized indirectly by instruments based upon the air-leakage principle (Kajanto *et al.*, 1998; Van Eperen, 1991), quite handy and fast for integration in production quality control schemes, but also somewhat uninformative regarding the nature of the irregularities that drive this phenomenon.

[32] A term related to the degree of uniformity in the fibre network that constitutes paper.

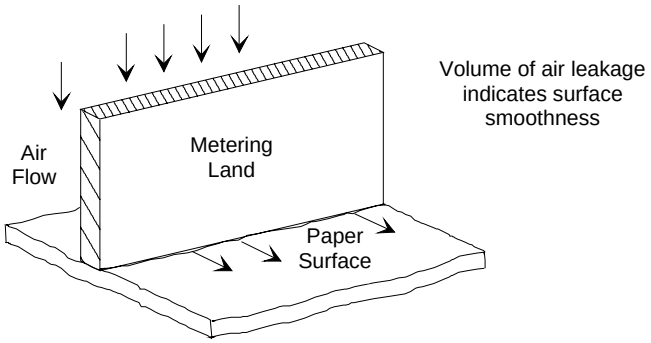


Figure 8.1. Schematic representation of the underlying measurement principle for common air-leakage equipments (Van Eperen, 1991).

Waviness, on the other hand, refers to those larger scale deviations from a flat shape, an example of which are the so called "piping streaks", that consist of streaks aligned along the largest dimension of paper, 1-3 *cm* wide, that may develop as a consequence of different fibre alignment streaks across the paper machine (Sze & Waech, 1997), but other representatives do exist as well, including the so called "flutes/fluting" in heavy ink coverage areas (Nordström *et al.*, 2002), and cockling, which consists of small "bumps", 5-50 *mm* in diameter, occurring at random positions in the paper sheet as a consequence of hygroexpansivity and structural unevenness of paper (Kajanto *et al.*, 1998; MacGregor, 2001). Quite often these larger scale waviness phenomena are assessed by trained operators through subjective classification schemes based upon sensorial analysis using several criteria defined *a priori* by a panel of experts, but efforts have also been carried out towards the development of more systematic and instrumental-based methodologies, namely using optical technology (Nordström *et al.*, 2002) and mechanical stylus profilometry (Costa *et al.*, 2004).

Profilometry, in particular, is a technique that collects a detailed profile of the paper surface. This raw profile can be processed afterwards, in order to calculate several parameters that summarize its main features at certain scales or frequency bands, where the analysis is to be focused. The complete surface profile contains all the raw information necessary to characterize the phenomena located at a relatively wide range of scales, ranging from a few micrometers to a few centimetres (Wågberg & Johansson, 2002).

Measurement procedure

The basic steps for the measurement procedure using profilometry (Costa *et al.*, 2005) appear illustrated in Figure 8.2. First, a sample of appropriate dimensions ($10\,cm \times 14\,cm$) is cut and positioned in the sheet support unit, especially designed for this application, so that the larger dimension of the item is perpendicular to the direction along which the measurement is to be carried out (usually paper cross-direction, CD, as the type of waviness phenomena we are most concerned with, i.e. "piping streaks", do always occur along a direction that is perpendicular to this one[33]). Then, a specified number of two-dimensional profiles are collected and the average values of the parameters recorded.

The measurement device used is a MahrSurf mechanical stylus profilometer, with a Perthometer S2 data processing unit, a drive unit PGK 120, and a MFW – skidless pick-up set. The profiles to be processed contain the central 6144 measures of surface height, separated by approximately $8.93\ \mu\text{m}$.

Parameters are computed internally in the data processing unit, after an intermediate step where roughness and waviness components of the original (de-trended) profile are separated. This separation is achieved by application of a digital filtering technique where a phase-corrected filter, with a selected cutoff frequency, is applied to the profile, in order to compute its components relative to roughness, containing the high frequency content of the profile, and that for waviness, relative to lower frequency oscillations. The value to be set for the cutoff is available in tables (also provided by the manufacturer of the equipment), according to certain geometrical characteristics of the surface, but irrespectively of the nature of the surface, be it a metal surface, stone or paper, for instance. In our application, the cutoff wavelength used was 2.5 *mm*. The raw profiles can also be saved for analysis, and, in fact, these will be used in the paper surface monitoring application to be addressed further ahead, in Section 8.2.4.

[33] The other directions of a sheet of paper are usually designated by "machine direction" or MD (i.e. the direction aligned with that regarding the paper sheet movement in the machine where it was produced) and thickness direction.

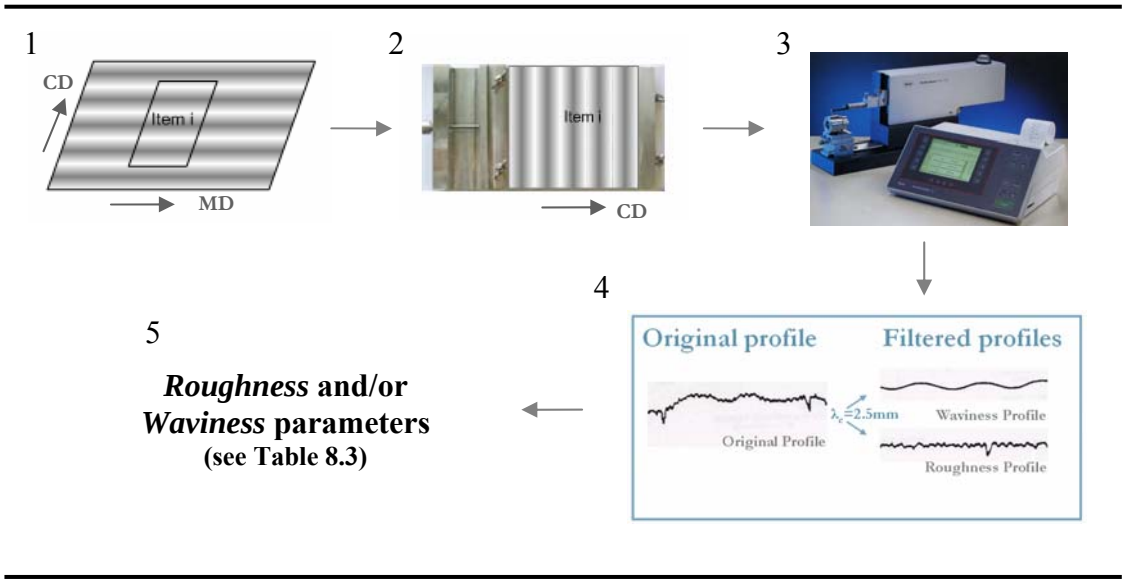


Figure 8.2. Steps involved in the measurement procedure using profilometry.

Parameters that can be computed in the data processing unit of the profilometer, summarizing the information contained in the entire waviness and roughness components, are presented in Table 8.3. For details regarding their precise mathematical definition, we refer to the relevant technical literature (ISO, 1996, 1997; Sander, 1991), as well as to the equipment's documentation.

Table 8.3. Waviness and roughness parameters obtained through profilometry.

Waviness	*Description*	*Roughness*	*Description*
Wt	*Total height of profile*	**Ra**	*Arithmetical mean deviation of profile*
Wa	*Arithmetical mean deviation of profile*	**Rz**	*Maximum height of profile*
W Sm	*Mean width of profile elements*	**Rq**	*RMS deviation of profile*
Wdq	*RMS slope of profile*	**Rp**	*Maximum profile peak height*
W	*Mean height of profile*	**Rt**	*Total height of profile*
W S	*Mean distance between local peaks*	**R S**	*Mean distance between local peaks*
AW	*Mean with of waviness motif*	**R Sm**	*Mean width of profile elements*
Wx	*Maximum height of waviness motif*	**R Sk**	*Skewness of profile*
Wte	*Total height of waviness motif*	**R Ku**	*Kurtosis of profile*
CMP	*Ratio between non combined roughness motifs and combined roughness motifs*	**Rv**	*Maximum profile valley depth*
Zeros	*Number of times that the profile crosses zero value*	**Rdq**	*RMS slope of profile*
P Sk	*Skewness of profile*		
P Ku	*Kurtosis of profile*		

8.2.2 Application of Profilometry to Predict Paper Surface Quality

In this sub-section, the built-in profilometer functionality for bi-scale profile analysis is explored. We will evaluate the potential of using information contained in the parameters relative to waviness and roughness phenomena (in this sub-section we will refer to such parameters simply as *features*) in order to predict quality assessments made by panels of experts.

The evaluations made by experts, regarding paper smoothness quality (related to roughness phenomena) and waviness, were carried out through sensorial analysis, and are supposed to be, therefore, close to what final consumers "fill" or "perceive" about product quality. Classification models developed can therefore provide an alternative way for assessing paper surface quality, which is faster, more objective and stable along time. The measurement scale used to classify the quality of paper samples is based on three levels: 1 – Bad, 2 – Moderate, 3 – Good.[34]

Since the problem consists of predicting the class membership of *objects* (samples) based on a set of *features* (i.e., a *classification* problem) and, furthermore, labelled information regarding a set of objects is available in the training stage, this problem falls under the broad scope of *supervised machine learning* methodologies. Therefore, several different approaches belonging to this class of methods were tested, in order to gain insight both into the prediction accuracy that can be achieved and the type of approaches that are more adequate for each situation. In particular, the following *classifiers* (supervised classification methods) were used (Fukunaga, 1990; Hastie *et al.*, 2001; Saraiva & Stephanopoulos, 1992; Theodoridis & Koutroumbas, 2003):

- Linear Classifier (Normal density-based; "Linear");
- Quadratic Classifier (Normal density-based; "Quadratic");
- Decision Tree Classifier ("Tree"); the classification tree algorithm implemented in CART® (Salford Systems) was also tested for the waviness classification problem;

[34] All data sets were collected in the context of a cooperation research project between several elements of the GEPSI research group and Portucel, SA.

- *k* – Nearest Neighbour Classifier ("kNN");
- Parzen Classifier ("Parzen");
- Neural Network Classifier (back-propagation; "NN back-prop.").

The following software was used to perform data analysis: Statistica© (Statsoft, Inc.), CART® (Salford Systems), PRTools4 (Duin *et al.*, 2004) and Matlab (The MathWorks, Inc.).

Predicting paper smoothness quality

The "paper smoothness" data set contains 22 features (11 roughness parameters for profiles taken along the MD and CD paper directions), and there are 36 labelled records available for training (6 from class 1, 18 from class 2, and 12 from class 3). Each datum in such a table is the result of averaging the parameter values obtained for three successive profiles taken over each sample.

An exploratory data analysis reveals the presence of a significant amount of correlation in this data set. This can be verified by examining its correlation map (a graphical representation of the correlation matrix, where correlation coefficients are coded as colours, Figure 8.3).

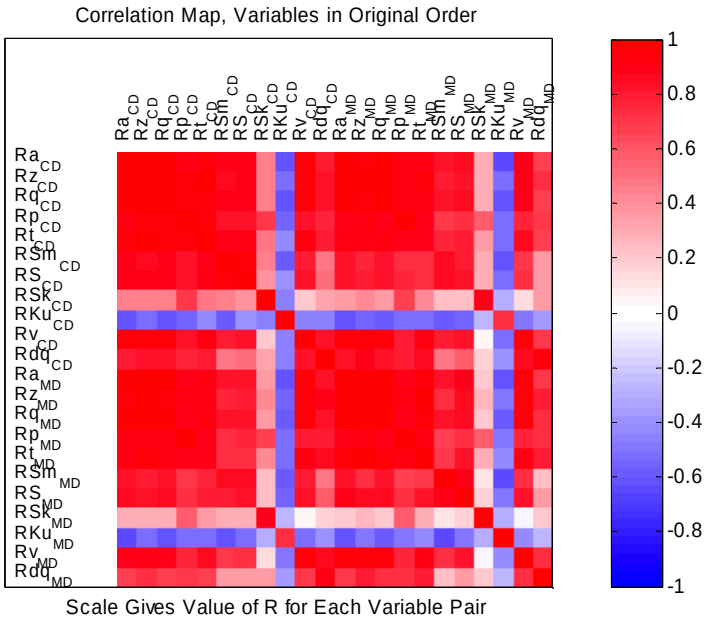


Figure 8.3. Correlation map for the features present in the paper "smoothness" data set.

This redundancy in the data set can also be checked through a "scree" plot (or eigenvalue plot), that represents the eigenvalues for the correlation matrix arranged in a decreasing order of magnitude (Figure 8.4), where it is also possible to see that a few number of components are dominating the entire data set variability.

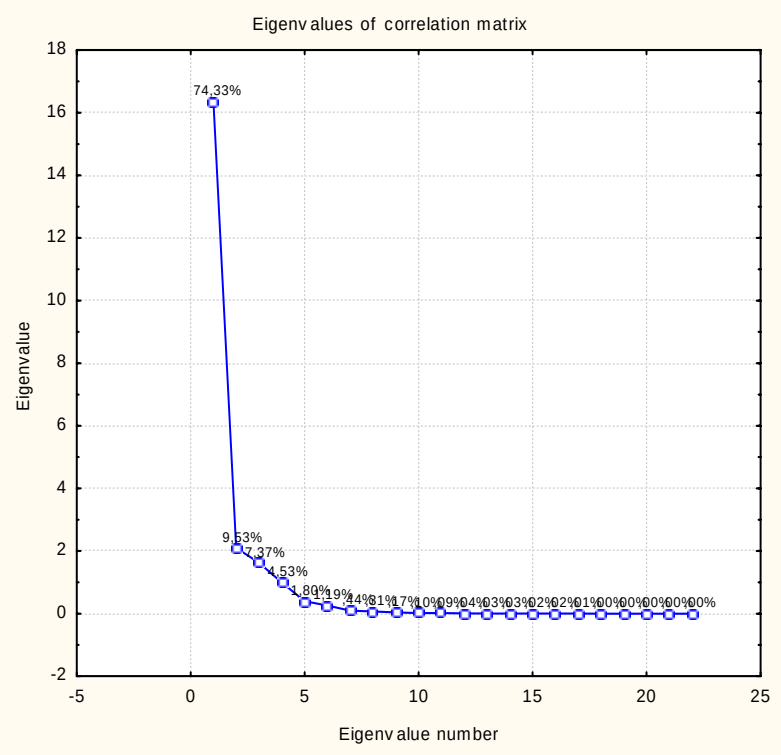


Figure 8.4. "Scree" plot for the "paper smoothness" data set.

The fact that most variables agglomerate after short clustering distances in an agglomerative clustering algorithm (Theodoridis & Koutroumbas, 2003) also means that they are quite similar (Figure 8.5-*a*). Furthermore, MD and CD parameter counterparts often provide identical roughness information, as they tend to agglomerate in pairs, something that can also be found out by analysing the loading plots for the first two principal components (Figure 8.5-*b*).

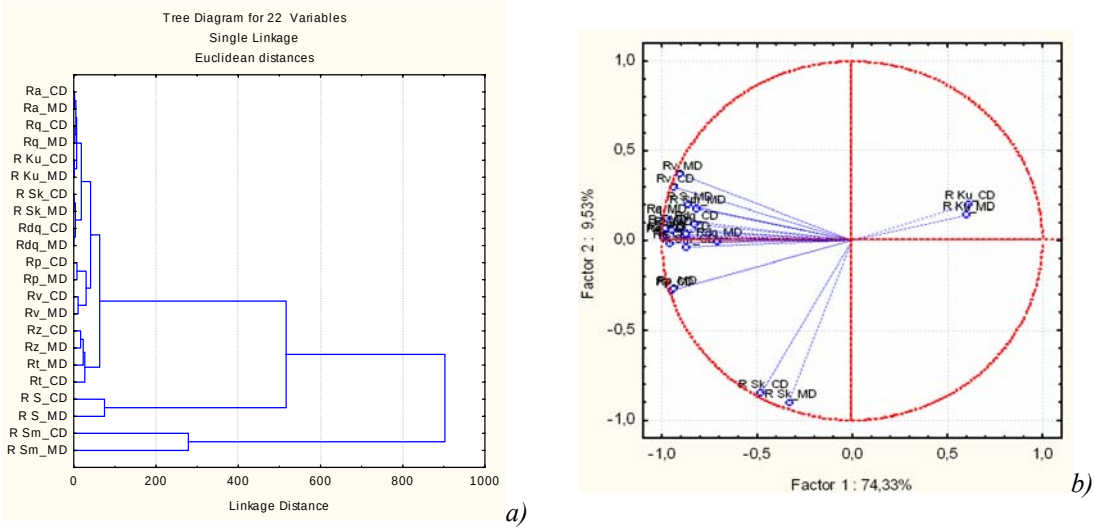


Figure 8.5. "Paper smoothness" data set: a) Tree diagram for the clustering of smoothness features (single linkage agglomerative algorithm using an Euclidean proximity measure); b) Loadings for the first two principal components.

Such a redundancy in the feature space not only means that not all of the features are bringing new relevant information for classification purposes, but it also often happens that using the full dimensional space in such circumstances may in fact be detrimental to classification performance and lead to unstable classifiers (Chiang *et al.*, 2001; Fukunaga, 1990; Naes *et al.*, 2002; Naes & Mevik, 2001; Theodoridis & Koutroumbas, 2003). Therefore, variables not bringing predictive power to the classification problem should be *removed* or *down-weighted*. In this study, different strategies were employed to reduce the effective dimension of the predictive feature space, i.e., the dimension of the subspace that is effectively used for classification purposes. These methodologies can be seen as *mappings* converting observations from the full feature space into another lower dimensional space, e.g., through projection-like operations or variable selection. The three different mappings strategies used are:

- Variable selection (VS);
- Principal Components Analysis (PCA);
- Fisher Discriminant Analysis (FDA).

The former technique maps the original feature space onto a subset preserving the most relevant variables for classification (in the sense of a given, previously defined, criterion). The second approach maps the features space by projecting it onto a lower dimensional one that retains most of the data variability. Finally, the last methodology also consists of projecting the original data into a lower dimensional plane, but this time estimated in order to provide maximal separation between classes.

Figure 8.6 illustrates the main steps involved in the implementation of classification procedures over new samples, after the classifiers and mappings have been estimated using training data with known class labels.

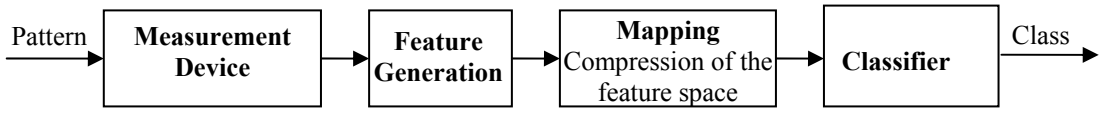


Figure 8.6. Main steps used in the implementation of classification procedures adopted in this study.

The results obtained for the paper "smoothness" data set are presented in Table 8.4. As the number of labelled objects available is not very large, the approach followed consists of estimating misclassification rates using a leave-one-out cross-validation procedure (LOO-CV). The results show that, in this case, the classification task is facilitated by the classes natural separation, and therefore some combinations of mapping/classifier, even though of a quite simple structure, such as VS/Tree or FDA/Linear, still work quite well (Figure 8.7). It is also apparent, from the results, that classifiers tend to perform better after an FDA transformation, followed by "variable selection" and finally by PCA.

Table 8.4. Misclassification rate estimates (LOO-CV) for the "paper smoothness" data set, using different combination of classifiers (first column) and mappings (first row).

↓ Classifier / Mapping →	VS	PCA	FDA
Linear	*0.1389*	*0.25*	*0*
Tree	*0*	*0.1667*	*0.0556*
kNN	*0.0278*	*0.0833*	*0*
Quadratic	*0.1389*	*0.1667*	*0*
Parzen	*0*	*0.0833*	*0*
NN (back-prop.)	*0.0556*	*0.25*	*0.0833*

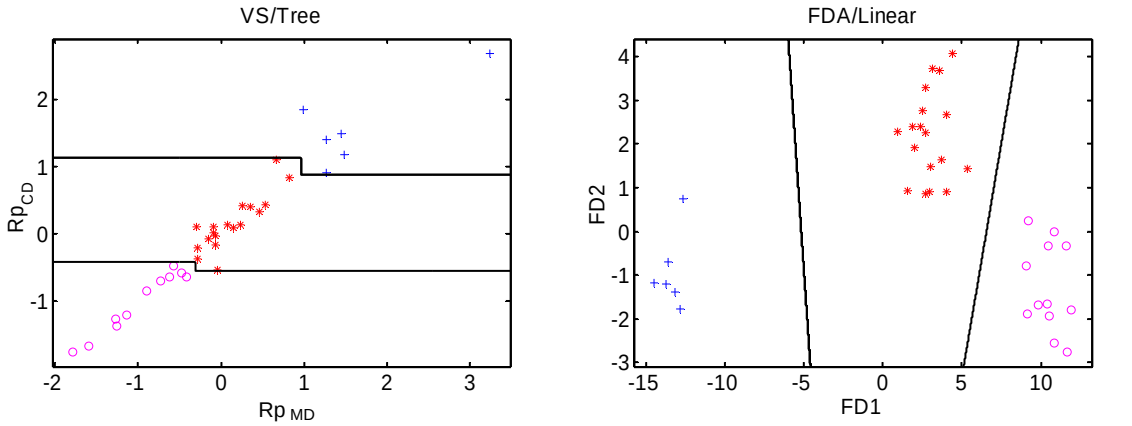


Figure 8.7. "Smoothness" data set: scatter plots with discriminant boundaries for the combination VS/Tree (a) and FDA/Linear (b).

Predicting paper waviness quality

Moving now to the "waviness" data set, it contains 13 features (waviness parameters for profiles taken in the paper CD direction, as the waviness phenomena we are mostly concerned with has its axis parallel to the MD direction), and 29 labelled records (9 from class 1, 12 from class 2 and 8 from class 3).

The degree of redundancy among features for this data set is less severe than what happened with the "smoothness" data set, as can be seen from its "scree" plot (Figure 8.8), where the eigenvalues magnitude difference is not as large, although some correlation is still present.

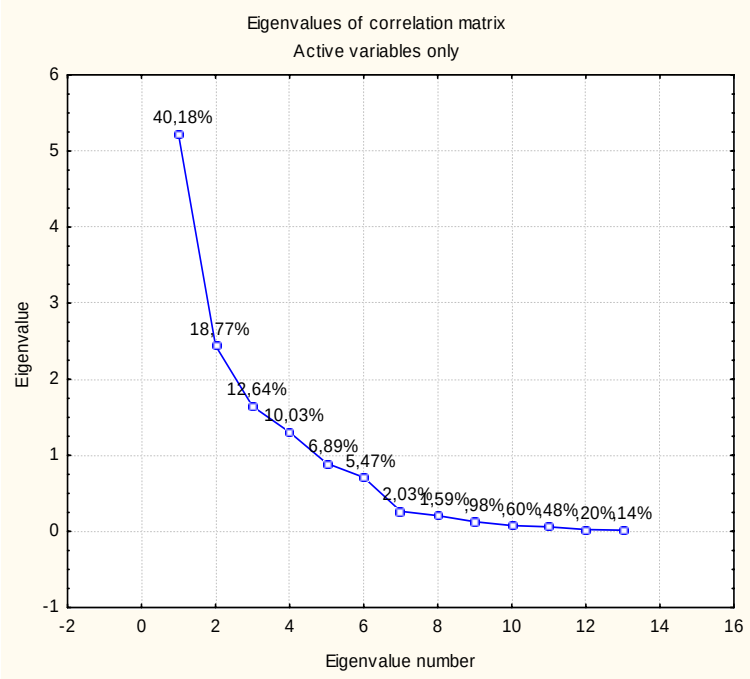


Figure 8.8. "Scree" plot for the "waviness" data set.

Analysing Figure 8.8, it is also possible to see that, for adequately describing the overall features variability, more than two dimensions should be retained, as they only explain a percentage of about $40.18\% + 18.77\% = 58.95\%$ of the overall variability (a preferable situation would be to retain 6 PC, for instance, which explain around 94% of variability). However, to enable a visual comparison of the results obtained for the different techniques, and explore the topology of the feature space, we keep the number of features to be used at two (as we did also for the "smoothness" data set).

Following a procedure similar to the one adopted for analysing the "smoothness" data set, the results presented in Table 8.5 were obtained.

The waviness data is more challenging from the standpoint of classification, being much more so when only two dimensional transformed feature spaces are used for classification. However, FDA does a quite good separation job (Figure 8.9), setting the ground for the achievement of interesting classification performances by some classifiers, such as Parzen or CART ® (that performed better than its implementation referred as "Tree", available in the "PRTools4" package). The VS and PCA methodologies would eventually require more dimensions in order to enable them to achieve a better separation of classes, but VS still performs better (similarly to what happened in the analysis of the "smoothness" data set).

Table 8.5. Misclassification rate estimates (LOO-CV) for the "waviness" data set, using different combination of classifiers (first column) and mappings (first row).

↓ Classifier / Mapping →	VS	PCA	FDA
Linear	*0.3448*	*0.3793*	*0.1724*
Tree	*0.2759*	*0.3793*	*0.1724*
kNN	*0.2069*	*0.5172*	*0.1379*
Quadratic	*0.3793*	*0.3793*	*0.2069*
Parzen	*0.3793*	*0.3793*	*0.0690*
NN (back-prop.)	*0.3103*	*0.5517*	*0.1724*

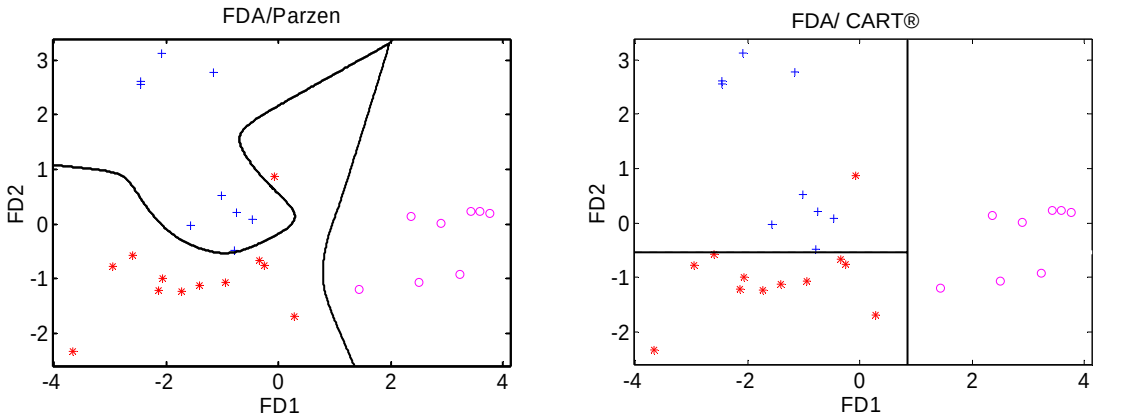


Figure 8.9. "Waviness" data set: scatter plots with discriminant boundaries for the combinations FDA/Parzen (a) and FDA/CART® (b).

Conclusions

In this sub-section, different supervised classification methodologies were applied to a "smoothness" data set and a "waviness" data set, containing labelled quality classes, in order to explore the possibility of developing predictive classification schemes, based upon profilometry measurements.

Results show that such classification tasks can be adequately addressed, even with a low dimensional predictive space (two dimensions).

The interesting results obtained for the classification tasks involving smoothness (related to roughness) and waviness phenomena, using the built-in bi-scale functionality of the profilometer, open good perspectives for a bi-scale monitoring scheme for the paper surface, based on raw profiles and a multiscale decomposition and analysis, made

independently of the measurement device's computations, and therefore more flexible in the sense that it can be better tailored to a particular situation.

8.2.3 Multiscale Analysis of the Paper Surface

In this section we look into the multiscale structure of paper surface, by analysing the *raw profiles* collected using profilometry. Our main goal here is to gain insight about the dominating surface phenomena at different scales and which scales can be attributed to each phenomenon. The knowledge gathered will then be used to set up a multiscale monitoring procedure for paper surface profiles, to be described in a subsequent sub-section of this chapter.

Graphical Representations of Multiscale Surface Phenomena

Let us perform a wavelet-based multiscale decomposition for a profile collected from paper exhibiting waviness phenomena. Figure 8.10 presents the details signals corresponding to several scales, as well as the approximation for the coarsest scale considered, along with their approximate wavelength ranges. These plots represent reconstructions in the original length domain of the events distributed across different frequency bands, according to the scale index. We also include information regarding the dominant surface phenomena at different scales, according to the literature (Kajanto *et al.*, 1998), and some accumulated engineering background knowledge about the subject. It is possible to detect an oscillation phenomenon characteristic of "piping streaks", on scales *10* and *11*, which, for the paper under analysis, typically occurs with a wavelength around $15\,mm$. However, by looking only at Figure 8.10, it is not straightforward to validate that scales relative to all roughness phenomena[35] for this particular type of paper are those proposed in the published literature. Therefore, to better discern where the transition between roughness phenomena and the next coarser scale phenomena really lies, other types of plots should be analysed, where the structure

[35] We analyze here the roughness phenomena all together.

of the surface across scales for the particular grade of paper under analysis is adequately summarized.

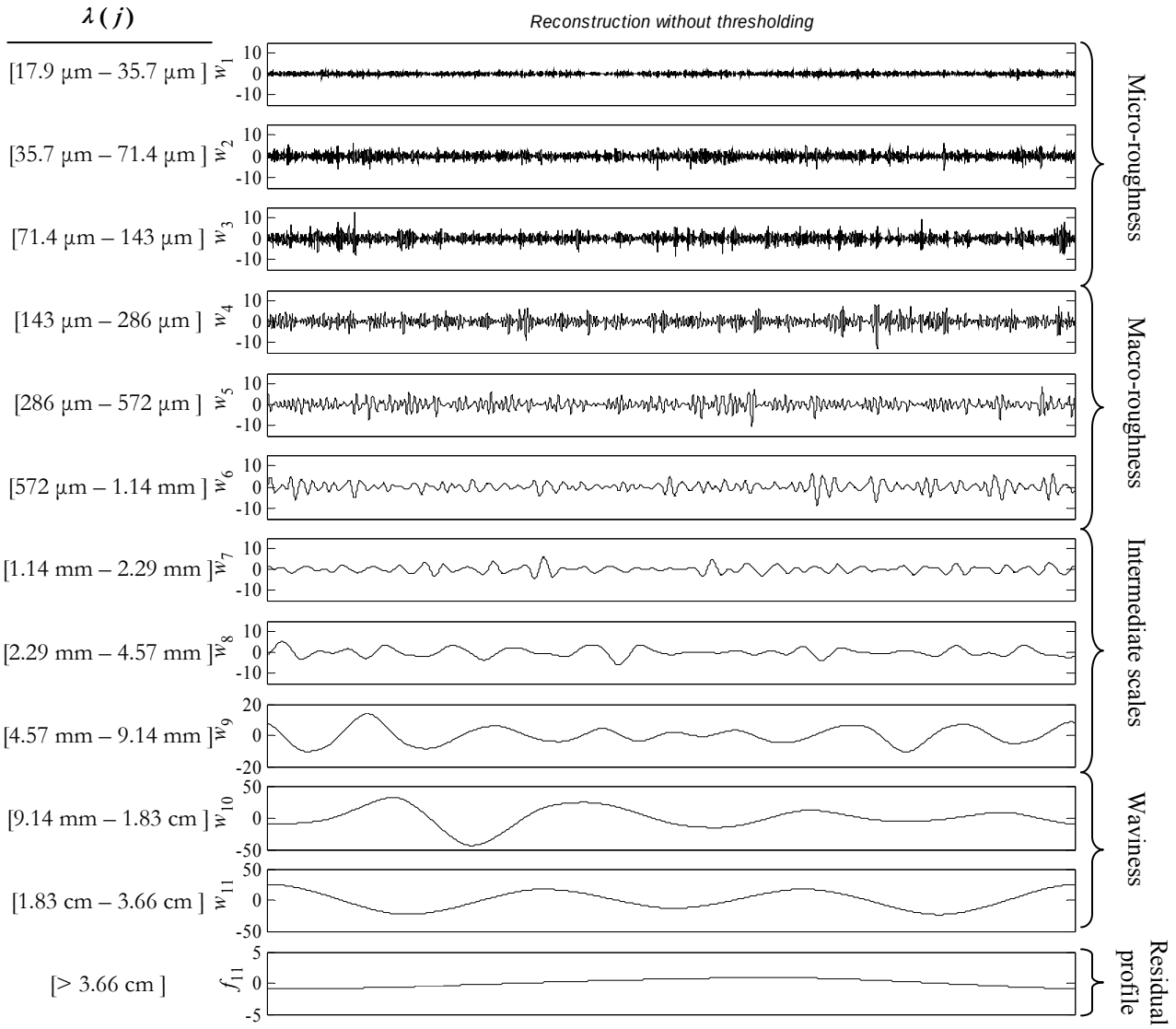


Figure 8.10. Plot of reconstructed detail signals at each scale ($w_j, j = 1:11$) along with the reconstructed approximation at the coarser scale considered, $j = 11$ (f_{11}). The approximate wavelength bands covered at each scale are presented on the left, and the designation of the surface phenomena relative to the scales presented, according to information available in the literature, are identified on the right.

One example of such a type of plot is presented in Figure 8.11, where the variance of detail coefficients at each scale is represented as a function of the scale index in a log-

log plot. We found out that, using this type of plot, the profiles for the analysed paper are such that there is a remarkable linear behaviour in the finest scales region, relative to roughness phenomena,[36] therefore facilitating the task of figuring out where the transition occurs by looking to the scale where such a smooth behaviour begins to break down. In the present case, it is possible to detect a change of pattern occurring slightly before scale 6, indicating that the roughness phenomena seem to collapse somewhere between scales 4 and 6, rather than between scales 6 and 7 (vertical lines), as suggested in the literature (also shown in Figure 8.10).

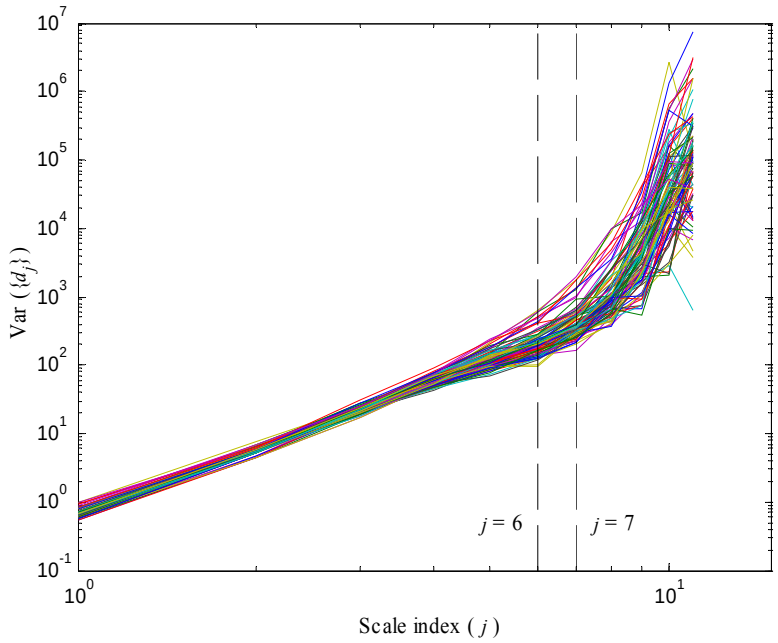


Figure 8.11. Log-log plot of the variance of detail coefficients at each scale (j), for 90 surface profiles taken in the paper cross direction. These samples have different levels of waviness magnitude, but similar roughness behaviour. Vertical lines indicate a transition region for the roughness phenomena, according to the literature (Kajanto *et al.*, 1998).

[36] This feature might very well be a "finger print" of the paper production process, that may be explored in other tasks as well, such as the prediction of some surface related quality parameters (e.g. friction or printability).

As opposed to the waviness phenomenon, which is usually easy to describe and identify as an oscillatory trend, the characterization of roughness phenomenon for a given paper may require a more involved analysis, given its stochastic behaviour. Therefore, the issue of portraying roughness phenomena for the paper grade under analysis was further pursued here, using a time series analysis approach.

Time Series Analysis of the Paper Surface's Roughness Phenomena

An adequate description of roughness for the paper grade under analysis must produce a power spectrum compatible with the results presented in Figure 8.11, which renders inadequate some descriptions from the field of statistical geometry for random fibre networks, leading to simple *iid* Normal or Poisson models with high mean values (Kajanto *et al.*, 1998), as they do not give rise to power spectra with such features.[37] On the other hand, analysing the surface height distributions in roughness profiles, we have often found distributions slightly skewed towards the left, which are also described by other authors (Forseth & Helle, 1996). Therefore, in order to develop a model for the (cross direction) roughness of the paper grade that we want to describe, an approach based on time series theory was adopted (Box *et al.*, 1994; Ljung, 1999), and a suitable autoregressive moving average model (ARMA) fitted to data.

In this context, an ARMA(2,2) was found to be the lowest order model that passes both residual autocorrelation (Figure 8.12) and partial-autocorrelation (Figure 8.13) validation analysis (Box *et al.*, 1994; Chatfield, 1989).

[37] In particular, they lack autocorrelation modelling, arising from the natural dependencies between measurements of surface height in adjacent positions.

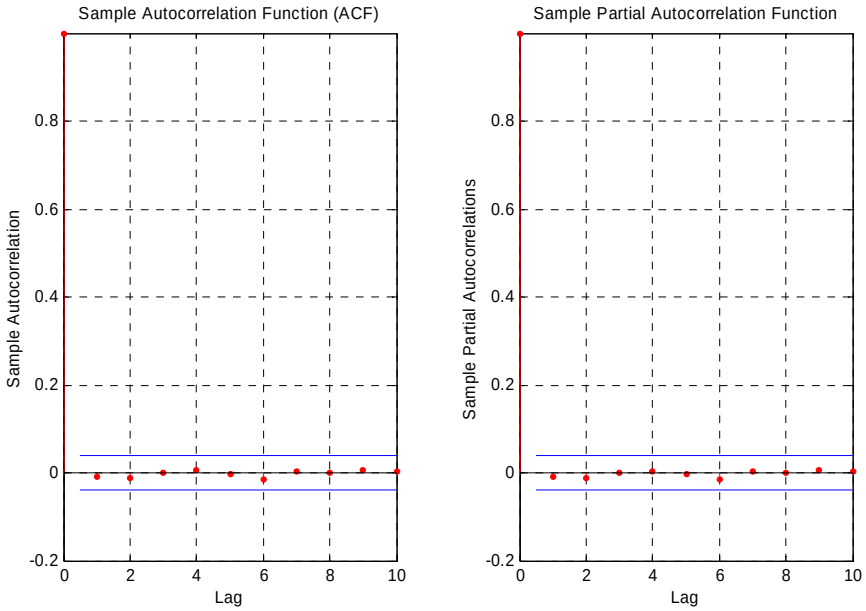


Figure 8.12. Sample autocorrelation and partial autocorrelation functions of the residuals obtained after adjusting an ARMA(2,2) model to a typical roughness profile. No significant autocorrelation structure is left to be explained in the residuals.

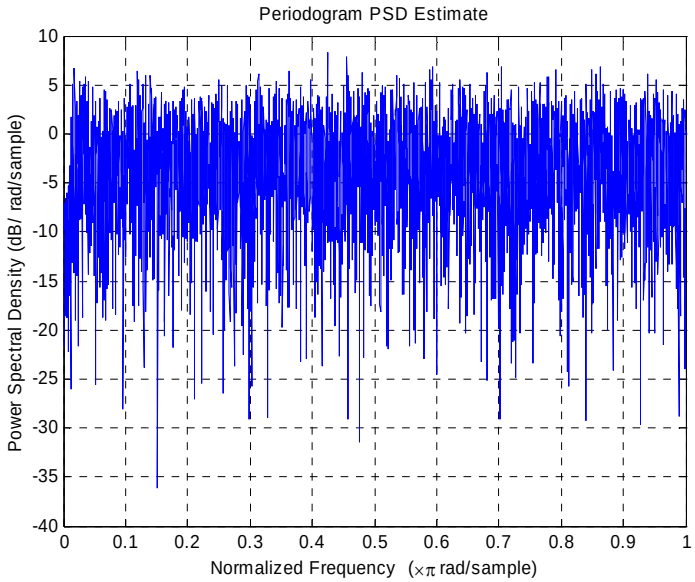


Figure 8.13. Power spectral density for the residuals obtained after adjusting an ARMA(2,2) model to a typical profile. Despite its "noisy" behaviour, the power spectrum mean level is fairly constant along the frequency bands, meaning that residuals behave like a random white noise sequence.

From all the normal operation roughness profiles, a typical one was chosen to fit the ARMA model parameters, thus leading to:

$$
\begin{aligned}
&A(q)W(t) = C(q)e(t), \text{ with} \\
&A(q) = 1 - 0.6605q^{-1} - 0.09479q^{-2} \\
&C(q) = 1 + 0.8111q^{-1} + 0.2365q^{-2} \\
&e(t) \sim iid\ N(0, \sigma_e^2),\ \sigma_e^2 = 2.3320
\end{aligned}
\tag{8.1}
$$

where $A(q)$ and $C(q)$ are polynomials in the shift operator, q, such that $q^{-1}W(k) = W(k-1)$, i.e., $W(t) + a_1 W(t-1) + a_2 W(t-2) = e(t) + c_1 e(t-1) + c_2 e(t-2)$, for an ARMA(2,2) model. Figure 8.14 illustrates the validity of the estimated model regarding a description of the true raw profile, in terms of the sample autocorrelation and partial-autocorrelation functions. It also reproduces the desired power spectrum behaviour within the roughness scales range.

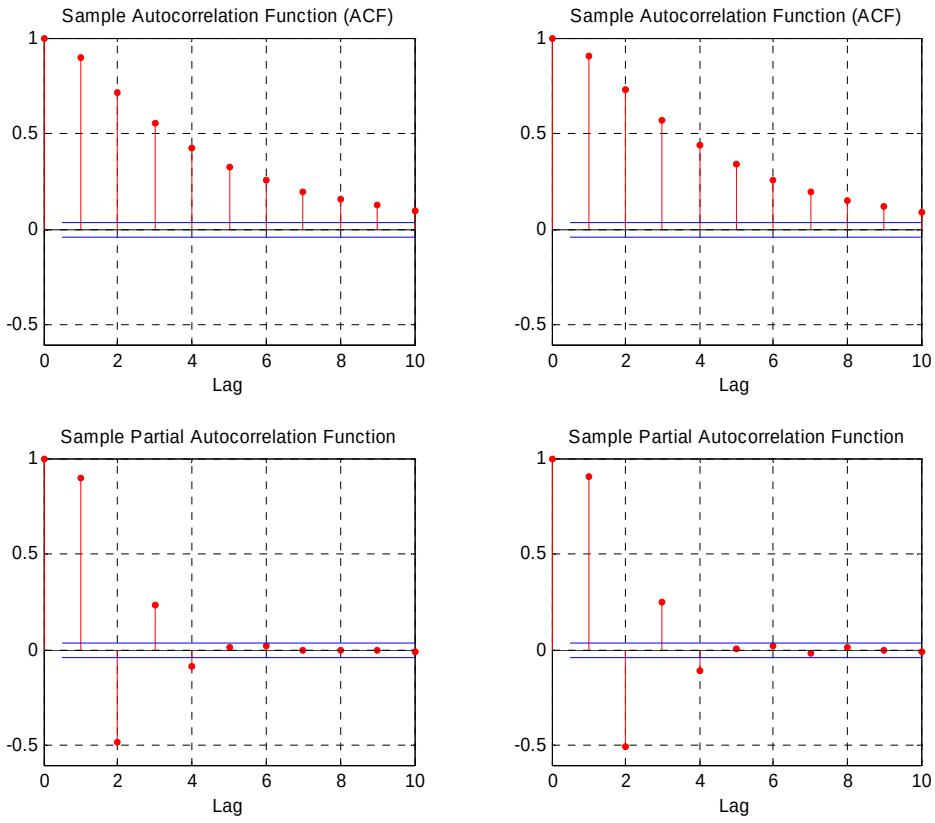


Figure 8.14. Sample autocorrelation and partial autocorrelation functions for a real roughness profile (left) and for a simulated profile using the estimated model, equation (8.1) (right).

To check that the proposed model is indeed representative of the roughness phenomena in the set containing all profiles analysed, Table 8.6 presents the means and standard deviation of the distribution of values for the parameters obtained by fitting an ARMA(2,2) model to each one of the 90 roughness profiles analyzed. As can be seen, the parameters for the model presented in (8.1) are quite typical of what can be found in this grade of paper, as they lie clearly inside the intervals established by the sample mean plus/minus one standard deviation.[38]

Table 8.6. Means and standard deviations for the ARMA(2,2) model parameters estimated using each one of the 90 profiles.

Parameter	*Mean*	*Standard deviation*
a_1	-0.6553	0.1772
a_2	-0.0799	0.1491
c_1	0.7618	0.1758
c_2	0.2086	0.0924
σ_e^2	2.4913	0.2343

8.2.4 Multiscale Monitoring of Paper Surface Profiles: Results

In this section the proposed multiscale monitoring procedure for stationary profiles is applied in the scope of the simultaneous monitoring of paper roughness and waviness phenomena. In this context, from all the scales available upon the wavelet decomposition of profiles, we will be only concerned with two sets of them: one set corresponding to roughness phenomena and another one to waviness phenomena. Several simulated scenarios, regarding paper surface, as well as real industrial data, are used for testing our methodology. In the simulation approach, realistic paper surface profiles are generated, representing a variety of situations that go from typical normal operating conditions to several degrees of abnormal situations (moderate and high), in order to evaluate the sensitivity of the proposed methodology to detect shifts, and

[38] Other approaches involving the use of time series analysis to characterize paper surface can be found elsewhere (Kapoor & Wu, 1978, 1979).

therefore its potential adequacy for real world implementations. Then, using real paper surface profiles, we tested how the methodology performs in practice through a set of approximately one hundred cross direction paper surface profiles, representing mainly different levels of waviness magnitude, but where some abnormal roughness behaviour can also be found.

Regarding the basic implementation steps mentioned in Table 8.1, the following paragraphs summarize relevant information for application to the present case study.

Acquisition of Profile

As already referred (Section 8.2.1), profiles are acquired along the CD paper direction, using a MahrSurf mechanical stylus profilometer set, with a Perthometer S2 data processing unit, a drive unit PGK 120, and a MFW – skidless pick-up set.

Wavelet Decomposition

The decomposition depth used in step 2 is $J_{dec} = 11$, so that the frequency ranges where "piping streaks" do develop can be adequately covered. An orthogonal Symmlet-8 wavelet filter (Mallat, 1998) was employed, because: 1) the shape of its associated wavelet does resemble that of waviness profiles; 2) it is smooth; 3) does have a compact support; and 4) is more symmetric (by design) then filters from the Daubechies orthogonal wavelet family.

Selection of Scales Relative to Each Phenomenon

Step 4 requires a preliminary selection of scales relative to roughness and waviness phenomena (conducted in step 3). As already mentioned, engineering knowledge refers that roughness scales range up to 1 *mm*, meaning that the maximum scale index should be somewhere between 6 and 7 (because, $10^{-3} m \in \left[2^6, 2^7\right] \times 8.93 \times 10^{-6} m$). On the other hand, by carefully analysing the multiscale patterns for different metrics, in several profiles with varying waviness magnitudes, but approximately the same roughness behaviour, and, in particular, if we analyse the variance of the detail coefficients at each scale (Figure 8.1), one can clearly detect a change of pattern occurring slightly before

scale 6, indicating that roughness phenomena stop somewhere between scales 4 and 6. Therefore, balancing these two pieces of information, the maximum scale index for roughness phenomena was set equal to 6, as an adequate compromise between engineering knowledge available and the analysis performed over a selected group of samples. The scale indices associated with these phenomena are then as follows: $J_{Rog} = \{1,2,...,6\}$. As to those relative to waviness, the maximum scale index is limited by the decomposition depth scale (11), and the minimum scale index was set equal to 10, in order to capture the minimum scale associated with "piping streaks" surface irregularities, since $2^{10} \times 8.93 \times 10^{-6} m \approx 0.01 m$ $(1 cm)$. Thus, the scale indices adopted for monitoring waviness phenomena are $J_{Wav} = \{10,11\}$.

Another task to be performed in step 4 regards the calculation of parameters that summarise relevant characteristics of the two phenomena, to be employed for statistical quality control purposes. Many metrics have been proposed to characterize both roughness (e.g. arithmetical mean deviation of profile, maximum height of profile, RMS deviation of profile, etc.) and waviness profiles (e.g. total height of profile, mean width of profile elements, slope of profile, etc.), that can be consulted in the profilometry literature (Sander, 1991) and norms (ISO, 1997), to which we can sum up others based upon wavelet coefficients (e.g. variance of detail coefficients distributed across selected scales for each phenomenon, and its slope in a log-log plot for roughness scales). As many of these metrics give rise to highly correlated data sets, when used together, we can either use them all and compress the monitoring dimension space, using, for instance, PCA, or choose a subset that provides all the important profile information for monitoring purposes, and set up control charts only for this subset. Using extended simulations and analysing real paper profiles, we found out that often a single, adequately chosen, parameter is good enough to detect magnitude changes in the roughness and waviness phenomena. This parsimonious solution works quite well, but can also be easily extended to incorporate more parameters. Therefore, the parameter (*statistic*, in the usual statistical terminology) chosen for monitoring roughness is the sample or empirical variance of the reconstructed roughness profile:

$$Empirical\ variance\ of\ roughness\ profiles = \frac{\sum_{k=1}^{N} \left(R_k - \overline{R} \right)^2}{N-1} \qquad (8.2)$$

where $R \equiv \{R_k\}_{k=1:N}$ is the roughness profile (N stands for the number of points in the roughness profile, which is also the same as the length of the original profile), obtained by performing an inverse wavelet transformation of the vector of wavelet coefficients where the only non-zero elements are those relative to the roughness scales, or, equivalently, $R = \sum_{i \in J_{Rog}} w_i$; $\overline{R}$ corresponds to its sample average (note that both the roughness profile, R, and the projections to the detail spaces, w_i, are vectors of the same dimension, N). As for the chosen waviness parameter (once again a statistic under the usual statistical terminology), we defined a simple magnitude parameter that correlated quite well with the visual assessment of waviness profiles, given by the maximum deviation from the mean value, $D_{\max}$, defined as:

$$D_{\max} = \max\left(C_p, C_v\right) \tag{8.3}$$

where C_p and C_v represent, respectively, the largest peak height and the largest valley depth of the profile centred at its mean value, $C(x) = W(x) - Z_m$, with Z_m defined as:

$$Z_m = \frac{1}{x_{\max} - x_{\min}} \int_{x_{\min}}^{x_{\max}} W(x)\, dx \tag{8.4}$$

i.e., $C_p = \max\left(C(x)\right)$ and $C_v = \min\left(C(x)\right)$ ($x_{\min}$ and $x_{\max}$ represent the initial and final X-axis coordinates, to be considered for the purpose of calculating Z_m); W is the waviness profile, obtained through the same procedure adopted for R, but using waviness scales instead in the reconstruction algorithm, $W = \sum_{i \in J_{Wav}} w_i$.

SPC Monitoring

The two parameters referred above are used to monitor multiscale phenomena in step 4, through two separate Shewhart control charts for individual observations (Montgomery, 2001). Their upper control limits were set through a non-parametric approach, using a Gaussian kernel density estimation methodology (Silverman, 1986) over reference data that correspond to normal operation conditions. As the underlying reference distribution depends strongly upon real industrial production conditions, and since no sufficient data are available at the moment to describe it thoroughly using a parametric approach, this alternative allowed us to assess the potential utility of our methodology. Furthermore, other SPC procedures can also be implemented in the future, such as CUSUM or EWMA, to enhance sensitivity to small shifts, as extensions of the proposed approach.

In step 5 we provide the operator with a diagnosis tool that maps each waviness profile into a two dimensional plot of $\lambda_{\max}$ versus $D_{\max}$, where $\lambda_{\max}$ stands for the finite wavelength where power spectra reach a maximum. Since "piping streaks" are well localized in the frequency domain (they have a characteristic wavelength typically somewhere around 20 *mm*, although this value depends upon a specific paper machine), such a plot allows for the fast identification of those high magnitude samples that may be classified into this type of abnormality. Several reference horizontal lines assist operators in the classification of the magnitude of the phenomena into three quality classes (good, intermediate, bad), which reflect the perception of a panel of experts, afterwards translated into values of $D_{\max}$. Another vertical reference line provides a separation between two wavelength ranges, one of which regards the "piping streaks" characteristic wavelength domain (Figure 8.18).

Simulation Results

Our simulation study provides an assessment of the underlying potential for the proposed methodology under simulated, though realistic, scenarios. As the behaviour of the true underlying industrial process, and therefore that of the monitoring statistics, are both rather complex and, to a larger extent, remain unknown at the present stage, the results presented here serve the purpose of evaluating its potential, deferring an accurate characterization of its Phase 2 performance (e.g., through ARL, ATS metrics) to future

work, when sound statistical modelling becomes possible with the availability of larger data sets.

To design a realistic simulation study, both waviness and especially roughness phenomena were carefully analysed, in order to estimate adequate models that are compatible with the main features present in real world paper surface profiles. Model (8.1) was used to generate the roughness component of the overall simulated profile (*R*). As for the waviness component (*W*), both the type of waveforms typically found when "piping streaks" are present, as well as other lower frequency irregularities and normal operation conditions profiles, were simulated through the superposition (sum) of several sinusoidal waveforms, $W = \sum_{i=1}^{n_W} W_i\left(\lambda_i, A_i\right)$, each one with its own wavelength (λ_i) and amplitude (A_i). We used four of such elementary waves ($n_W = 4$) to synthesize the overall waviness profiles, through the sequence of steps presented in Table 8.7.

Table 8.7. Sequence of steps involved in the generation of the waviness component for the overall profile.

1. Definition of simulation parameters, including average wavelength ($\overline{\lambda}$), wavelength half range ($\Delta\lambda$), average maximum amplitude ($\overline{A}_{\max}$) and amplitude range ($\Delta A_{\max}$);
2. Generate wavelengths λ_i for each component wave, W_i, where $\lambda_i \sim U\left(\overline{\lambda} - \Delta\lambda, \overline{\lambda} + \Delta\lambda\right)$, *i*=1:4, with $U(\cdot)$ representing an uniform distribution in the range specified as argument;
3. Generate amplitude $A_{\max}$ for the final (overall) waveform W, where $A_{\max} \sim U\left(\overline{A}_{\max} - \Delta\overline{A}_{\max}, \overline{A}_{\max} + \Delta\overline{A}_{\max}\right)$;
4. Definition of amplitudes for each component wave, W_i, calculating first the unscaled amplitude for each component, A_i^*, and then scaling the four components in order to obtain a final waveform with the amplitude specified in step 3, i.e. $A_i^* \sim U\left(\overline{A}_{\max} - \Delta\overline{A}_{\max}, \overline{A}_{\max} + \Delta\overline{A}_{\max}\right)$, $A_i = A_i^* A_{\max} \Big/ \sum A_i^*$;

5. Generation of individual wave components with the parameters computed in the previous steps and using the same sampling spacing and number of points as for the real profiles (8.93 μm and 6144, respectively); perform summation to obtain the resulting waviness profile, $W = \sum_{i=1}^{4} W_i\left(\lambda_i, A_i\right)$.

Finally, both the roughness and waviness profiles are combined to obtain the simulated raw profiles, P ($P = R + W$). The proposed approach was tested under several scenarios, in order to assess its potential to detect shifts of different magnitude in the waviness profile, as well as shifts in roughness. Figure 8.15 presents the MS-SPC control charts for data generated according to the five simulation scenarios described in Table 8.8.

Table 8.8. Simulation parameters associated with different scenarios studied.

↓ **Scenario / Simulation parameter** →	$\overline{\lambda}$ (***mm***)	$\Delta\lambda$ (***mm***)	$\overline{A}_{\max}$ (***μm***)	$\Delta A_{\max}$ (***μm***)	**Roughness model**
1. Normal operation	40	10	30	20	(3.4)
2. "Piping streaks", moderate magnitude	17	3	70	20	(3.4)
3. "Piping streaks", high magnitude	17	3	110	20	(3.4)
4. "Cockling", high magnitude	80	20	100	20	(3.4)
5. Roughness, high magnitude	40	10	30	20	(3.5)

The first two plots (a and b) refer to control charts for roughness and waviness, respectively, with 99% control limits established after a preliminary Gaussian kernel density estimation step, where 40 samples representing normal operation conditions were used, whereas plot c) combines them into a single plot.[39] The non-parametric estimation approach was adopted, in order to overcome the difficulties raised by the shapes of the distributions found for the monitoring statistics, which do not resemble any known probability density function. Under such circumstances, the Gaussian kernel

[39] Lines in this plot are control limits for each parameter, represented only for reference, not aiming to define the combined 99% control region, although this could also be done within the scope of non-parametric approaches (Martin & Morris, 1996).

density method provides an adequate way to estimate the underlying distribution, through an adequate fit/smoothness trade-off (Silverman, 1986).

From Figure 8.15 (a and b), we can verify that all the shifts simulated under conditions 2-5 are clearly detected in the appropriate control chart, even the one for moderate "piping streaks" irregularity. In Figure 8.15-c, one can notice an overlap occurring in the region of significant waviness phenomena, where "piping streaks" of different magnitude and "cockling" appear superimposed. However, since the former has a quite localized behaviour in the frequency domain, these two types of phenomena can be quite well resolved under the current simulation conditions, by bringing in an extra classifying element, which is the (finite) wavelength where the waviness profile power spectra reaches its maximum, $\lambda_{\max}$.

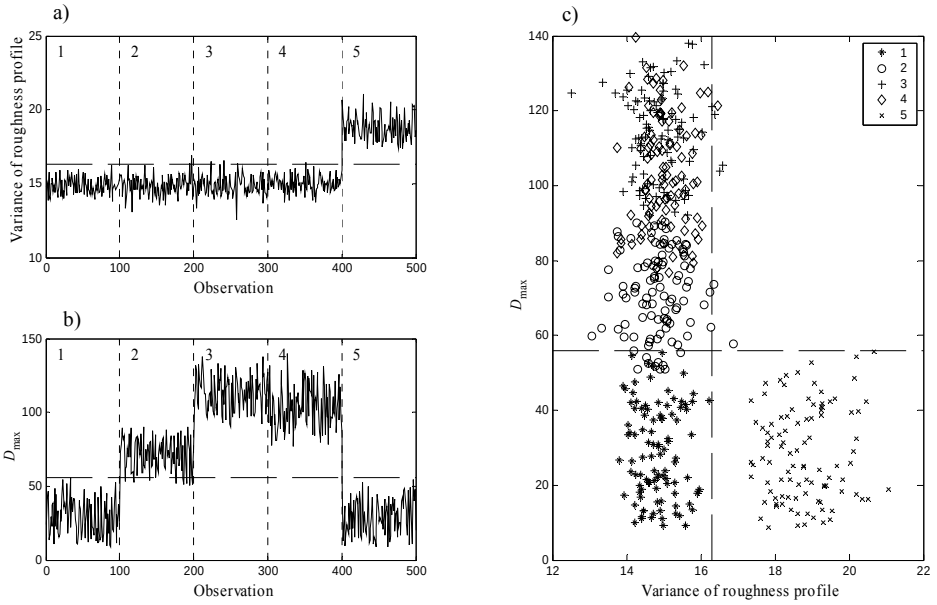


Figure 8.15. Control charts for monitoring roughness (a) and waviness (b), both with 99% upper control limits, and a combined plot that monitors both statistics (c). The five sectors indicated in plots a) and b) and the symbols used in plot c) refer to the simulation scenarios described in Table 8.8.

Figure 8.16 presents such a plot, where we can see that a separation is indeed possible between these two phenomena (Figure 8.15-c is the orthogonal projection of the points in this three-dimensional plot, onto the "variance of roughness profile" *versus* "$D_{\max}$" plane). Since we are particularly concerned with following "piping streaks", this idea

will be pursued a bit further in the next subsection, in order to develop a plot that indicates when such phenomena may be occurring.

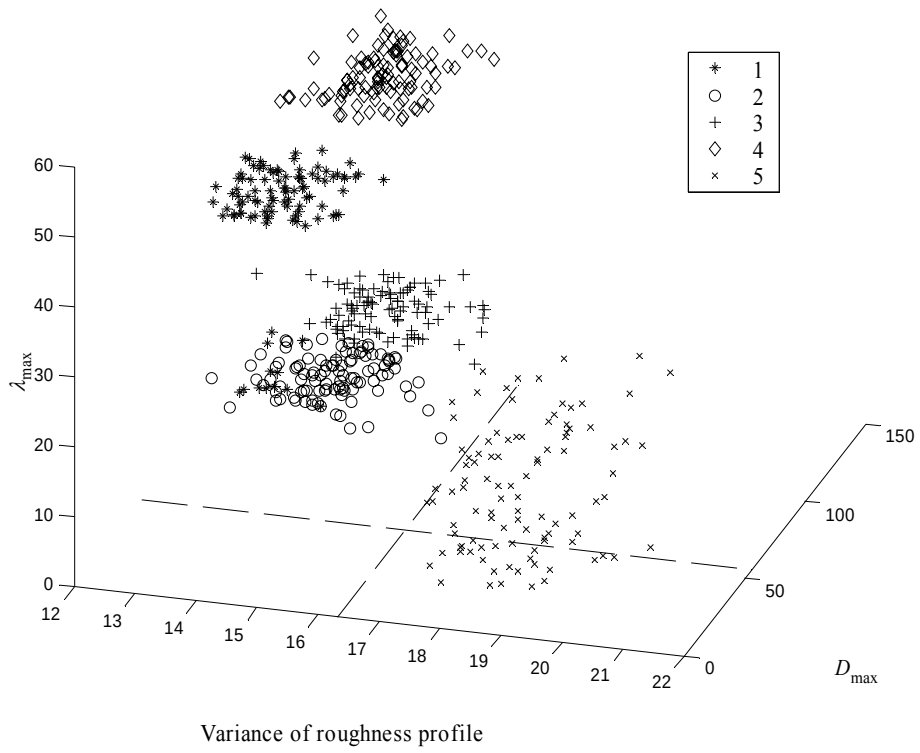


Figure 8.16. A three dimensional plot of the variance of roughness profiles *versus* D_{max} and λ_{max}. Symbols refer to the scenarios described in Table 8.8. Waviness (2-3) and cockling (4) clusters appear now quite well separated.

Multiscale Monitoring of Real Paper Surface Profiles Results

To further test the multiscale profile monitoring approach under conditions even closer to those found in real industrial practice, a pilot study was run in the context of a collaboration between our research group and Portucel (a major Portuguese pulp and paper company). Approximately one hundred profiles were gathered, containing samples within the normal operation quality standards, as well as others corresponding to several types of abnormal situations. Table 8.9 presents a general description of the samples whose profiles were used in this study.

Table 8.9. Description of surface phenomena exhibited by real surface profiles.

Description	***Samples***
Reference set	1-40
No waviness	41-61
Moderate waviness	62-82
High waviness	83-88
Upward trend on Bendtsen roughness	89-98

Control limits were set based on the variability exhibited by the samples from the reference set, following the same approach used with simulated data. The test set contains samples with low, moderate and high waviness, as well as samples that correspond to an upward trend in roughness magnitude, as measured by the Bendtsen tester (Kajanto *et al.*, 1998; Van Eperen, 1991), an instrument based on the air-leakage principle that measures the volume of air flowing between a ring and the paper surface. As no roughness measurements were available for the former samples, with various levels of waviness magnitude, it is not possible to analyse the monitoring performance of the roughness chart for such samples. Some moderate and high waviness samples can be classified into typical "piping-streaks" and "cockling" representatives by looking at their profiles, but for others that is not possible. We will refer to them simply as (high or moderate) waviness samples.

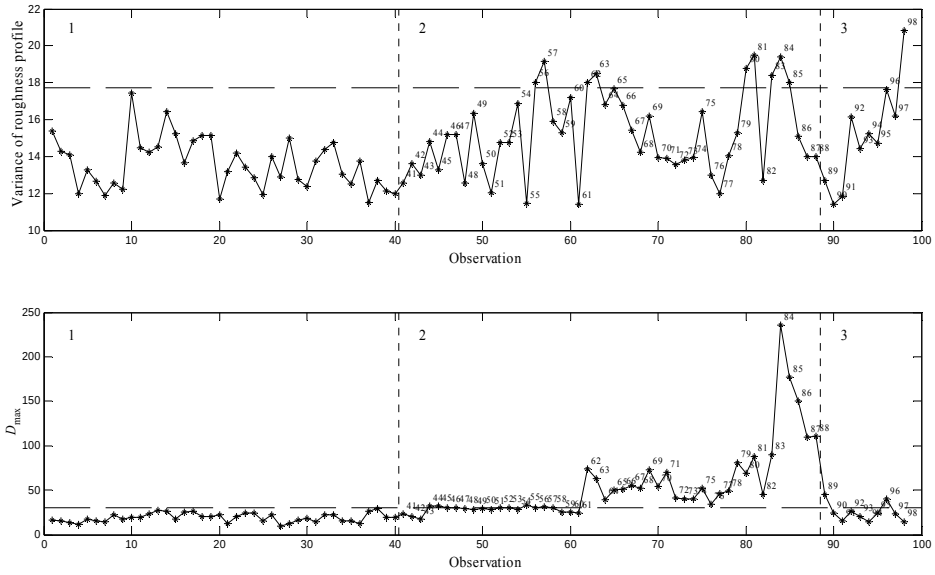


Figure 8.17. Control charts for monitoring roughness (a) and waviness (b). The first part of the data sets (1) regards reference data, the second (2) is relative to waviness phenomena with different magnitudes (see Table 8.9 for details) and the third (3) regards an upward trend in roughness, as measured by the Bendtsen tester.

Figure 8.17 presents our MS-SPC monitoring results for the real profiles. We can see that the SPC chart for monitoring waviness does indeed follow the magnitude trends of the samples described in Table 8.9. As for the chart relative to roughness, it is also possible to verify that it captures the upward trend in the last 10 samples, besides other significant events scattered over other samples in the test set. To facilitate the detection of samples with "piping streaks" waviness, a two-dimensional plot of $\lambda_{\max}$ *versus* $D_{\max}$, presented in Figure 8.18, was adopted, where the samples appear segregated along the vertical direction according to the magnitude of the waviness phenomena, and along the horizontal direction, according to their characteristic wavelength. In general, this plot enables a correct separation, especially when samples present well defined waviness behaviour, such as is usually the case when "piping streaks" occur. The horizontal classification boundaries, presented in Figure 8.18, were set by analysing the location and localization of the samples classified into three waviness magnitudes classes, through a simple procedure that weights the natural upper and lower boundaries for each adjacent class, using the number of elements in each class, whereas the vertical classification line was drawn using engineering knowledge regarding "piping streaks".

From what was presented in these two studies, we can see that the proposed multiscale profile monitoring methodology can indeed be used for monitoring simultaneously both paper waviness and roughness phenomena.

8.3 Conclusions

In this chapter, a multiscale profile monitoring approach was presented, discussed and applied to the simultaneous monitoring of both roughness and waviness paper surface phenomena in an integrated way. Its monitoring performance was analysed through simulated realistic scenarios and using real industrial data. The approach is built around a wavelet based multiscale decomposition framework, that essentially conducts a multiscale filtering of the raw profile, effectively separating the two phenomena under analysis, making also use of available engineering knowledge and information derived from an analysis of the distributions of different quantities through the scales.

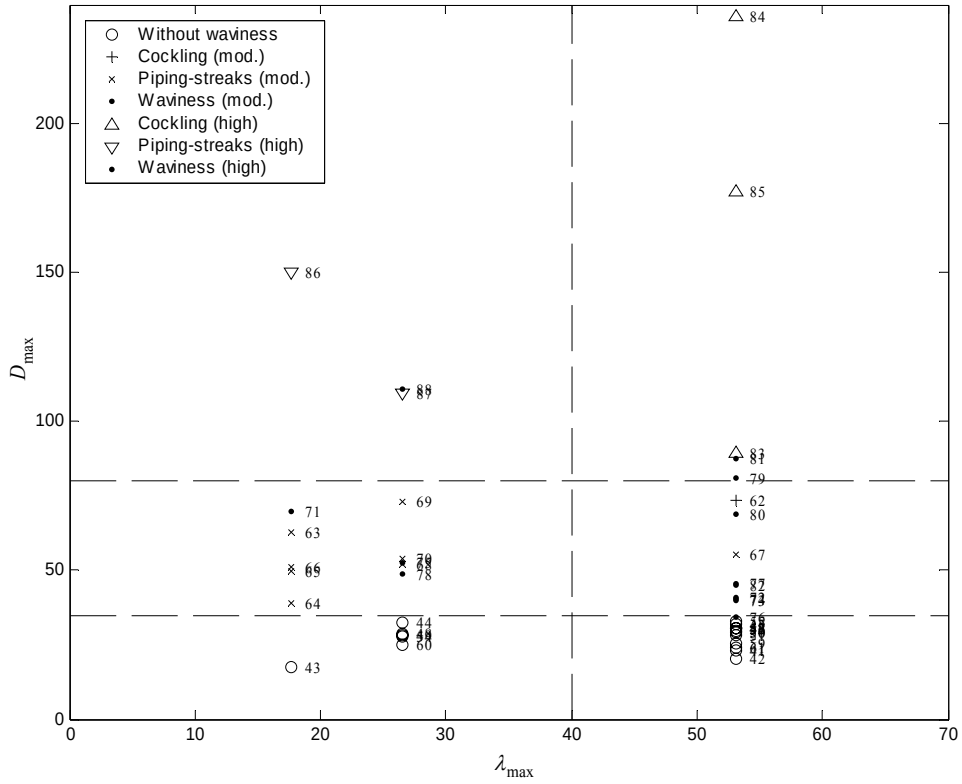


Figure 8.18. Plot of λ_{max} *versus* D_{max} for the real profiles data set. In this plot, waviness phenomena are classified into three levels of magnitude, separated by horizontal lines (low at the bottom, moderate at the middle and high at the top), and in two regions of characteristic wavelength, the range at the left being characteristic of "piping streaks" phenomena.

The results presented for the case study related with monitoring of paper surface using profilometry allow us to conclude in favour of the adequacy of adopting the proposed approach for monitoring simultaneously both types of phenomena (roughness and waviness), but its thorough characterization in terms of Phase 2 detection performance metrics (ARL, ATS) is deferred until more process data can be accumulated.

Chapter 9. Multiscale Statistical Process Control with Multiresolution Data

In this chapter, we focus on monitoring time domain phenomena, where a multitude of features can develop, making wavelet-based multiscale approaches adequate in this context, given their well known feature extraction effectiveness. In particular, an approach for conducting multiscale statistical process control that adequately integrates data at different resolutions (multiresolution data), called MR-MSSPC, is presented. Its general structure is based on Bakshi's (1998) MSSPC framework, designed to handle data at a single resolution. Significant modifications were introduced in order to process multiresolution information. The main MR-MSSPC features are illustrated through three examples, and issues related to real world implementations and with the interpretation of the multiscale covariance structure, are addressed in a fourth example, where a CSTR system under feedback control is simulated. The proposed approach proved to be able to provide a clearer definition of regions where significant events occur and a more sensitive response when the process is brought back to normal operation, when compared to approaches based on single resolution data.

9.1 Introduction

Data generated in chemical process plants arise from many sources, such as on-line and off-line process sensors, laboratorial tests, readings made by operators or raw materials

specifications, to name just a few. To such a variety of origins are usually associated complex data structures, due to diversity in time acquisition, missing data patterns, as well as in variable resolutions, since the values from different variables may carry information that covers different time ranges (multiresolution data). In spite of several developments have been proposed to address the sparsity problem created by multirate (Izadi *et al.*, 2005; Lu *et al.*, 2004; Tangirala, 2001) and missing data (Arteaga & Ferrer, 2002; Little & Rubin, 2002; Nelson *et al.*, 1996; Walczak & Massart, 2001), the issue of handling multiresolution process data remains, to a large extent, unexplored, with developments mainly centred around signal and image processing problems (Bassevile *et al.*, 1992a; Chou *et al.*, 1994a; Willsky, 2002).

In a multiresolution data structure, we can find variables whose values are collected punctually (high time resolution) at every node of a fine grid whose spacing is established by their (also higher) acquisition rates, and variables that represent averages over larger time ranges (i.e. over several nodes of this grid), to which we will refer to as "lower resolution variables" (the term averaging support, *AS*, will also be used to address the period of time, or number of nodes, over which averages are computed).

In industrial applications, multiresolution data structures usually arise when process sensor information is combined with data taken from other sources, such as the following: averages made by operators from several readings taken from process measurement devices during their shifts, which are then annotated in daily operation reports or introduced manually in a computer connected to the central data storage unit; measurements from pools of raw material or products accumulated during a period of time and mixed thoroughly before testing; averages of process variables taken over a period of time, which are computed automatically by local DCS computers (e.g. on an hourly basis); aggregated measurements from each batch operation.

On the other hand, processes going on in chemical plants are themselves typically quite complex, and this complexity is also reflected in collected data, which contain the cumulative effect of many underlying phenomena and disturbances, with different location and localization patterns in the time/frequency plane. Not only the overall *system* has a multiscale nature, since it is composed of processing units that span different time scales and frequency bands, but also the *inputs* (manipulation actions, disturbances, faults) can present a variety of features with distinct time/frequency characteristics. For such reasons, multiscale approaches designed to handle and take

advantage of the information contained at different scales have been developed for addressing different tasks (Bakshi, 1999; Motard & Joseph, 1994), namely process monitoring.

In this context, multiscale monitoring approaches provide an adequate basis for developing a (multiscale) process monitoring framework that integrates information with different resolutions, as the concept of resolution (or scale) is already present in their algorithmic structures, by design, in particular for those based on the wavelet transform as a tool for separating dynamic features, contained at different scales. Therefore, the structure underlying Bakshi's (1998) MSSPC (for data at a single resolution, Section 2.5.1) was adopted in this work as an adequate starting point to integrate data with different resolutions, a research topic also covered by a number of authors, as referred in Section 2.5.1.

The remaining parts of this chapter are organized as follows. In the next section, MSSPC is reviewed, now focusing on some important implementation details rather than the more generic presentation of Section 2.5.1. In the third section, the proposed MSSPC approach that integrates multiresolution data (MR-MSSPC) is introduced. Then, in the following section, several examples illustrate the improved effectiveness achieved with our methodology, in identifying the region in the time domain where a fault occurs and its promptness in detecting transition points, when compared with other alternatives, based on single resolution data structures. A last example addresses the case of monitoring a non-linear multivariate dynamic process using MR-MSSPC, where several important practical issues, regarding its real world implementation, are referred, as well as some extensions, namely the possible definition of an adequate resolution for each variable being monitored. A final section summarizes the main results presented and conclusions reached.

9.2 MSSPC: Implementation Details

As already referred in Section 2.5.1., MSSPC is based on multiscale principal components analysis (MSPCA), which combines the decorrelation ability of PCA, regarding cross-correlations among variables, with that of the wavelet transform for any potential autocorrelated behaviour in each variable, and, furthermore the deterministic/stochastic separation power associated with this type of transform

(Bakshi, 1998). In summary, the MSSPC procedure consists of computing independent principal components models and control limits for SPC-PCA control charts, at each scale, using data collected from the process operating under normal conditions. Then, as new data is acquired, the wavelet coefficients are calculated at each scale, according to the chosen discretization procedure (to be described in the next section), and control chart procedures implemented separately at each scale. If any significant activity is detected at any scale, the signal is reconstructed back to the time domain, using only coefficients from the significant scales, a task that can be interpreted as a multiscale feature extraction mechanism. The covariance matrix at the finest scale is also computed, using information related with the significant scales, in order to implement the statistical tests at the finest scale (T^2 and Q control charts), that will produce the final outcome of the MSSPC method, stating whether the process can be considered to be operating under normal conditions, or if a special event has occurred.

Despite the well established methodological structure underlying MSSPC, described above, there are still some degrees of freedom left open on *how* certain tasks can be implemented, leading to different "flavours" regarding its exact algorithmic implementation. For instance, looking to the reconstruction stage, where the wavelet coefficients at significant scales are collected to reconstruct the signal in the original domain (finest scale), certain decisions have to be made, in order to answer questions such as:

- Should we use in the reconstruction the raw coefficients or their projections onto the PCA models at each scale? (Rosen, 2001);
- Should the projected data onto a PCA model at the finest scale (necessary to calculate the monitoring statistics) be obtained directly from the projections at each scale, or from a projection made at the finest scale, using the reconstructed data and a PCA model calculated from the combined covariance matrix?
- Regarding the way this combined covariance matrix is obtained, can we adopt, instead of the 1/0 weighting scheme proposed by Bakshi (1998), an alternative strategy that weights scales according to their relevance from the stand point of the events to be detected, in order to increase detection sensitivity and focus the method in the correct frequency range, tailoring it, by this way, to better suit its final monitoring goals? (Rosen, 2001)

Other questions arise at other levels and also need to be answered:

- During the start-up of the MSSPC methodology, before the full decomposition depth is attained since not enough data were collected yet, should the information regarding finer approximation coefficients be used for monitoring?
- Should an equal number of principal components be adopted at all scales (Bakshi, 1998) or not (Rosen, 2001)?
- If a different number of components is used in the PCA models at each scale, which criteria should be adopted to set the number of components after reconstruction, when coefficients arise from scales where the PCA models have a different number of components?
- Should we scale the original data, i.e., before applying the wavelet transform, or the wavelet coefficients at each scale?

All these decisions influence the final MSSPC algorithm adopted, and therefore one should be aware of them when comparing different MSSPC approaches. However, overall performance is not expected to vary quite significantly, as all alternatives share the same basic structure, that being the key factor contributing to the success of MSSPC methodologies.

9.3 Description of the MSSPC Framework for Handling Multiresolution Data (MR-MSSPC)

9.3.1 Discretization Strategies

Besides all the degrees of freedom mentioned in the previous section, another differentiating feature regarding MSSPC implementations, and an important one, regards the type of data windows over which the wavelet transformed is applied, and based upon which the subsequent analysis is carried out.

In one extreme, we have the moving window of constant dyadic length used by Bakshi (1998), that consists of translating a time window with length $2^{J_{dec}}$ (where J_{dec} is the decomposition depth of the wavelet transform used in the multiscale analysis), so that the last vector of observations is always included in the window, after an initial phase

that goes from observation number 1 to observation number $2^{J_{dec}}$, where its length increases in such a way that a dyadic window is always used, until it reaches the maximum length (Figure 9.1 I-a and II-a). This window can be used in the implementation of a fully on-line MSSPC procedure.

Concerning now procedures that involve a time delay, we have the moving window of variable dyadic length, that enables the successive calculation of the coefficients regarding an orthonormal wavelet transformation (Figure 9.1 I-b and II-b), in opposition to the coefficients of its undecimated counterpart, also known as translation invariant wavelet transform, that are calculated using the first type of moving window referred above. The former procedure corresponds to a uniform discretization of the wavelet translation parameter, while the latter implements a dyadic discretization strategy (Aradhye *et al.*, 2003). As can be seen from Figure 9.1 I-b and II-b, the length of the window is not constant along time, and therefore not all the wavelet coefficients are used for monitoring at each stage.

Finally, we have the non-overlapping moving windows of constant dyadic length ($2^{J_{dec}}$), over which all the relevant orthogonal wavelet coefficients can be calculated using batches of collected data, every $\left(2^{J_{dec}}\right)^{th}$ observation (Figure 9.1 I,II-c). This strategy also corresponds to a dyadic discretization of the wavelet translation parameter, but now all the coefficients for the selected decomposition depth (regarding a given data window) are calculated simultaneously, and not sequentially, as happens with the previous approach.

Let us now consider the situation where, among the collected data set, there are variables whose values regard averages over different time supports (multiresolution data). These values become available at the end of these periods, when they are recorded in the data storing unit. The traditional way for incorporating them in the monitoring procedures designed to analyze data at a single resolution usually consists of holding the last available value constant during the time steps corresponding to the finest resolution, when no new information is collected, until new average values become available (zero-order hold).

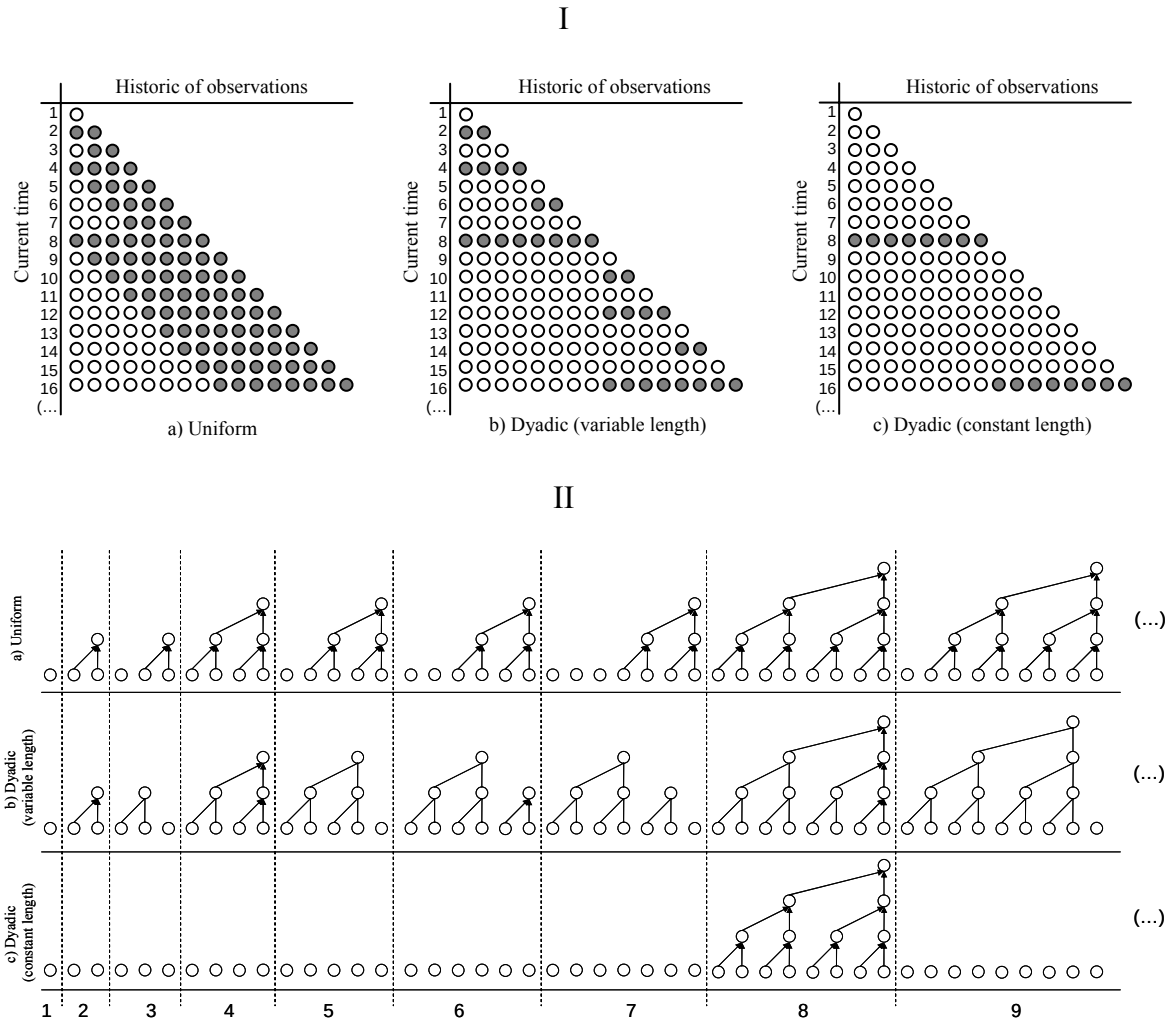


Figure 9.1. Two representations that illustrate different discretization strategies used in MSSPC, for $J_{dec} = 3$. Representation I illustrates which data points are involved in each window considered in the computations. Dark circles represent the values analysed at each time, which is represented in the vertical axis. The horizontal axis accumulates all the collected observations until the current time is reached (shown in the vertical axis). Representation II schematically represents the calculations performed under each type of discretization. The discretization methodologies considered are: a) overlapping moving windows of constant dyadic length (uniform discretization); b) dyadic moving windows for orthogonal wavelet transform calculations (variable window length dyadic discretization); c) non-overlapping moving windows of constant dyadic length (constant window length dyadic discretization).

This strategy creates a mismatch between the time support where the averages were calculated, and the one attributed to the average values. To illustrate this point, let us consider a situation where a variable corresponding to averages over four successive observations at the finest resolution is being acquired. Figure 9.2-a) illustrates the time ranges across which average values were calculated, while Figure 9.2-b) depicts the

ranges where the values are held constant with such a procedure. These are quite different, and in fact have only one interception point in the discretization grid at the finest resolution.

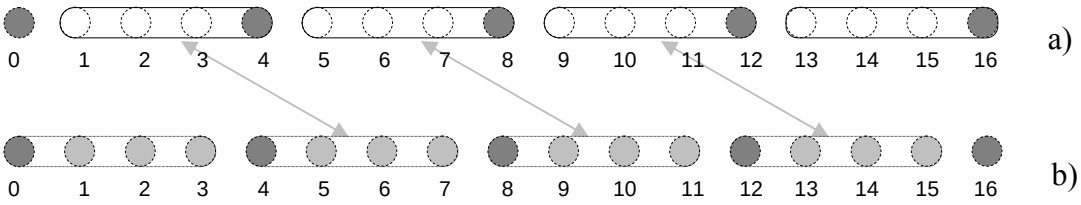


Figure 9.2. Time ranges over which average values are actually calculated (a) and those where the values are held constant in a conventional strategy to incorporate multiresolution data in single resolution methodologies (b).

From the different discretization approaches described above, the one that was found to be more adequate for setting up a multiresolution MSSPC procedure (MR-MSSPC) is the variable window length, dyadic discretization. As happens with its constant window length dyadic counterpart, this strategy has the important property of allowing for low resolution measurements to maintain their effective time supports (as represented in Figure 9.2-a), but without introducing as much time delay in the monitoring procedure. The uniform procedure was designed to handle on-line MSSPC tasks in situations where all variables have the same resolution (single resolution data). It is quite effective in such a context, but requires, for this same reason, a data pre-processing stage of the type represented in Figure 9.2-b.

9.3.2 Description of the MR-MSSPC methodology

The MR-MSSPC methodology begins with a specification of the resolution associated with the values collected for each variable. Quite often there is a finest resolution, corresponding to variables collected at higher sampling rates, which is used to set the finest grid (scale index $j = 0$). If variable X_i correspond to averages computed over time supports of length 2^{J_i} times that of the finest resolution, then its scale index, or resolution level, is set to J_i. A variable at resolution J_i can only be decomposed to

scales coarser (i.e., higher) than J_i, and therefore it does not contribute to the monitoring procedures implemented at finer scales ($j \le J_i$). This attribution is straightforward in situations where the low resolution variables represent averages over dyadic supports. In case this does not happen, we propose setting J_i as the immediately next coarser scale, i.e. $J_i = \lceil \log_2(AS) \rceil$, where AS is the averaging time support, $\lceil \log_2(AS) \rceil$ standing for the smallest integer n such that $n \ge \log_2(AS)$, and project data onto this scale using a weighted averaging procedure that will be described further ahead in this chapter.

The decomposition depth to be used in the wavelet transformation phase of standard MSSPC, J_{dec}, is another parameter to be set before implementation of the methodology. It must be higher than $\max\{J_i\}_{i=1:m}$ (usually $J_{dec} \ge \max\{J_i\}_{i=1:m} + 2$ for reasons related to the ability for reconstructing behaviour of past events). A summary of the whole procedure is presented in Table 9.1.

Table 9.1. Summary of MR-MSSPC methodology.

I. *Compute PCA models at each scale using reference data.*

a. For each variable ($X_i, i = 1:m$), perform the wavelet decomposition from $J_i + 1$ to J_{dec};

b. Calculate the mean vectors and covariance matrices at each scale;

c. Select the number of PCs and calculate PCA models at each scale.

II. *Implement MR-MSSPC methodology.*

a. For each observation index, *k*, multiple of $2^{J_{min}+1}$ ($J_{min} = \min\{J_i\}_{i=1:m}$);

i. Get dyadic window corresponding to current observation (length equal to $2^{J_{max}(k)}$);

ii. Decompose those variables X_i for which $J_i < J_{max}(k)$, from $J_i + 1$ to $J_{max}(k)$;

iii. Implement PCA-based MSPC at each scale where coefficients are available, using Hotelling's T^2 and Q statistics, and select the scales where significant events are detected from the standpoint of these

statistics (note that detail coefficients for some scales may not be available, while the scale of approximation coefficients to be used depends exclusively upon $J_{max}(k)$);

iv. Using the scales where significant events are detected, reconstruct data at the resolution levels between the coarser scale where a significant event was detected, J^*, and J_{min}, and that corresponding to the resolution for some variable, i.e. scales j that satisfy the condition $j : J_{min} < j < J^* \wedge j \in \{J_i\}_{i=1:m}$. Do the same for the mean vectors and covariance matrices associated with the selected scales.

v. Using the reconstructed data, recombined covariance matrices and mean vectors, calculate T^2 and Q statistics at each intermediate resolution, and look for significant events in these charts. If none detects a significant event, consider the process to be operating under normal conditions; if any of them shows an abnormal value, than trigger an alarm, and study the contribution plots for the reconstructed statistics at the scale where the signal occurs. The plots of the tests performed at each scale also contain information about the frequency ranges involved in the perturbation, and can be checked at a second stage of troubleshooting.

The PCA models developed in the initial stage, involving wavelet coefficients calculated from reference data (I), are not only for the detail coefficients at each scale ($0 < j \leq J_{dec}$) and for the approximation coefficients at scale J_{dec}, as happens with MSSPC with uniform discretization, but also for the approximation coefficients at scales $0 < j < J_{dec}$. This is due to the variable length associated with the type of windows used, which implies that very often the maximum decomposition depth is lower than J_{dec} ($J_{max} \leq J_{dec}$, where J_{max} is the maximum possible decomposition depth for the current data window). Thus, we often do have available approximation coefficients for $j < J_{dec}$, and, in practice, we found out that they actually can play an

important role in the earlier detection of sustained shits. Therefore, we include them in the implementation of MR-MSSPC.

Furthermore, PCA models at different scales may have a different number of variables associated with them, the models for the finest scales having fewer variables than those for the coarser scales, as these also integrate lower resolution variables. For this reason, the number of components for the models at each scale must be chosen individually. A rule has also to be defined in order to specify the number of PCs to use in the PCA model for reconstructed data at the finest scale, when it has contributions from several scales, whose models have different numbers of components. We will use the minimum, from all the scales involved in the reconstruction, so that the number of PCs will always be smaller than the number of reconstructed variables.

Stage II.iv corresponds to an extension of the MSSPC's reconstruction phase, using information from scales where significant events were detected, back to the original time domain, when multiresolution data are present. As we are now dealing with variables having different resolutions, we test the statistics derived from the reconstructed data at these intermediate resolutions, besides $j = 0$ (now converted to the more general J_{min}), provided that they stay below the coarser scale where significant activity was detected (if not, i.e, if the coarsest significant scale lies below the resolution of a given variable, then such a behaviour can not be due to its intervention, and therefore the reconstruction at those resolutions is not relevant). Therefore, we may end up with more than one plot for the reconstructed T^2 and Q statistics (one per resolution satisfying the conditions mentioned). Thus, in order to maintain the overall significance level of the SPC procedure adopted in the confirmation phase for each of the two statistics, control limits are adjusted using a correction factor applied over the significance level (α), derived from the Bonferroni inequality: α / n_{charts}, where n_{charts} is the number of charts used simultaneously for each statistic.

Another relevant issue regards the wavelet decomposition of variables available at coarser resolutions. Filtering operations, followed by dyadic down-sampling at each stage of the wavelet decomposition, encompass scaling operations that assure energy conservation for the orthonormal transformation (Parseval relation). As the coarser resolution variables have fewer decomposition stages than the other finer resolution variables, scaling operations that might have been made initially to the whole data set

would be now distorted at each scale, if no additional scaling is imposed to the coarser resolution variables. This scaling strictly depends on the difference between the finest resolution index and the resolution index for such variables, and it was implemented in the proposed methodology.

9.4 Illustrative Examples of MR-MSSPC Application

In this section, the main features of the proposed MR-MSSPC methodology are illustrated through its application to several different examples. The good properties of MSSPC methodologies in the monitoring of systems exhibiting autocorrelation were already widely explored in the literature (Aradhye *et al.*, 2003; Bakshi, 1998; Kano *et al.*, 2002; Misra *et al.*, 2002; Rosen & Lennox, 2001), and such properties are inherited by the proposed MR-MSSPC. In fact, since the methodology is based upon a dyadic discretization strategy, the decorrelation ability of the multiresolution decomposition is even higher than that obtained with an uniform discretization scheme. It is therefore expected to be even more suited to address highly correlated and nonstationary processes (Aradhye *et al.*, 2003). Thus, our focus in these studies is mainly over stationary uncorrelated systems, where the main features of the method can be more clearly illustrated, but an example is also presented regarding a more complex dynamic system (CSTR under feedback control), where several interesting features connected to real world implementations of the methodology are addressed.

The following latent variable model was adopted for data generation in the first three studies presented below (Bakshi, 1998), since this kind of model structure is quite representative of data collected from many real world industrial processes (Burnham *et al.*, 1999; MacGregor & Kourti, 1998):

$$X(k) = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & -1 \end{bmatrix} \cdot L(k) + \varepsilon(k) \tag{9.1}$$

$$L(k) \sim iid\ N(0, \Sigma_l),\ \Sigma_l = \mathbf{I}_2$$
$$\varepsilon(k) \sim iid\ N(0, \Sigma_\varepsilon),\ \Sigma_l = 0.2 \cdot \mathbf{I}_4$$

where $\mathbf{I}_m$ is the identity matrix with dimension *m*.

For the purposes of illustrating the MR-MSSPC framework, variable X_4 contains coarser resolution information, given as the successive averages over non-overlapping windows of length *AS* (to be defined for each example), while the remaining variables are all available at the finest resolution. Therefore, variable X_4 will only be acquired at the end of each period of *AS* consecutive observations, representing the mean of its values over that period of time.

In the first example, presented next, we illustrate the situation where the averaging window length, *AS*, is a dyadic number (2^{J_i}), leaving for the third example, a situation where such a support is non-dyadic.

9.4.1 Example 1: MR-MSSPC for Multiresolution Data with Dyadic Supports

A reference set with 4096 observations was generated using latent variable model (9.1). Variables $\{X_1, X_2, X_3\}$ are available at the finest scale ($J_1 = J_2 = J_3 = 0$), while variable X_4 represents averages over windows with length 4 ($J_4 = 2$). To test the monitoring features of MR-MSSPC, 128 observations are generated and a shift of magnitude +1 is imposed in all variables between observations 43 and 83 (included). The fact that transition times do not fall in the dyadic grid at a boundary between two averaging windows is intentional, in order to see how the method behaves in such less favourable conditions.

Figure 9.3 presents the results obtained regarding control charts for the T^2 and Q statistics at the two resolutions available in the data set, i.e. at $J_i = \{0, 2\}$. In the MR-MSSPC charts, circles (O) are used to indicate that the respective statistic's abscissa corresponds to the time where the last value of the method's analysing dyadic window

is acquired, and where calculations are actually made and results plotted.[40] For instance, in Figure 9.3, circles appear every 2nd observation in the plots for $J_i = 0$, and every 4th observation in the plots for $J_i = 2$, as only at such time instants new values become available for graphical representation at these resolutions. Therefore, a *decision about the state of the process is only taken at times corresponding to observations signalled with circles*. In case a statistic (T^2 or Q) signalled with a circle falls above control limits, its observation number also appears next to it. In such cases, all the values of this statistic regarding the same dyadic window are represented (×), as well as the associated control limits (–). This plotting feature is important to enable a more accurate reconstruction of the time at which a special cause occurred, even if it is detected at a later stage.

It was also decided to represent the points of the statistics and control limits even if the last observation is not significant, provided that there is at least one scale where a significant event was detected (with no number associated with it in the plots). This enables us to see more clearly when the process returns back to normal operation, as well as to visualize imminent abnormal situations in their early stages, when some unusual patterns become noticeable, prior to their full manifestation.

When no significant event is detected at any scale, a "zero" point is plotted (•).

From Figure 9.3, we can see that Q charts are more sensitive to the type of fault analysed in this example than T^2 charts. The Q statistic at $J_i = 0$ clearly indicates that an abnormal observation has occurred in the immediate past neighbourhood of observation 44, and that the process has returned back to normal shortly after observation 80. A mild spurious observation is also detected again at time 88, but the plot reconstructs quite clearly that the process has returned to normality.

[40] There is some delay during which data is collected and stored; thus, some observations are only plotted after some time, not corresponding to "current values", and therefore not being signalled with a circle.

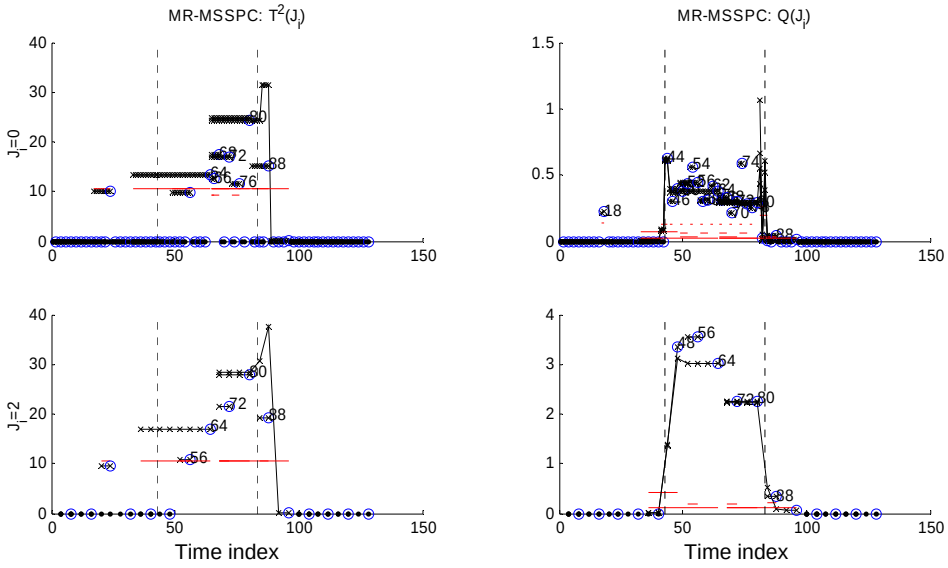


Figure 9.3. Plots of the T^2 and Q statistics at the two resolutions available in the data set, $J_i = \{0, 2\}$, using data reconstructed from significant scales. Control limits are set for a confidence level of 99% (horizontal line segments). Legend: ○ - signals effective plotting times ("current times"); × - appears if the statistic is significant at "current time", in which case its values in the same dyadic window are also represented (the "current time" index also appears next to the corresponding circle); – - control limit for the statistic, which is represented every time a significant event is detected at some scale relevant for the control chart; • - indicates a "common cause" observation (not statistically significant).

Figure 9.4 illustrates the underlying MR-MSSPC monitoring tasks conducted at each scale on the detail coefficients for $0 < j \le J_{dec}$ and on the approximation coefficients at scale J_{dec}, while Figure 9.5 regards the ones involving approximation coefficients for scales $0 < j < J_{dec}$. As can been seen from these plots, detail coefficients play an important role in the detection of transition times, while approximation coefficients have the complementary role of signalling abnormalities during the duration of a sustained shift.

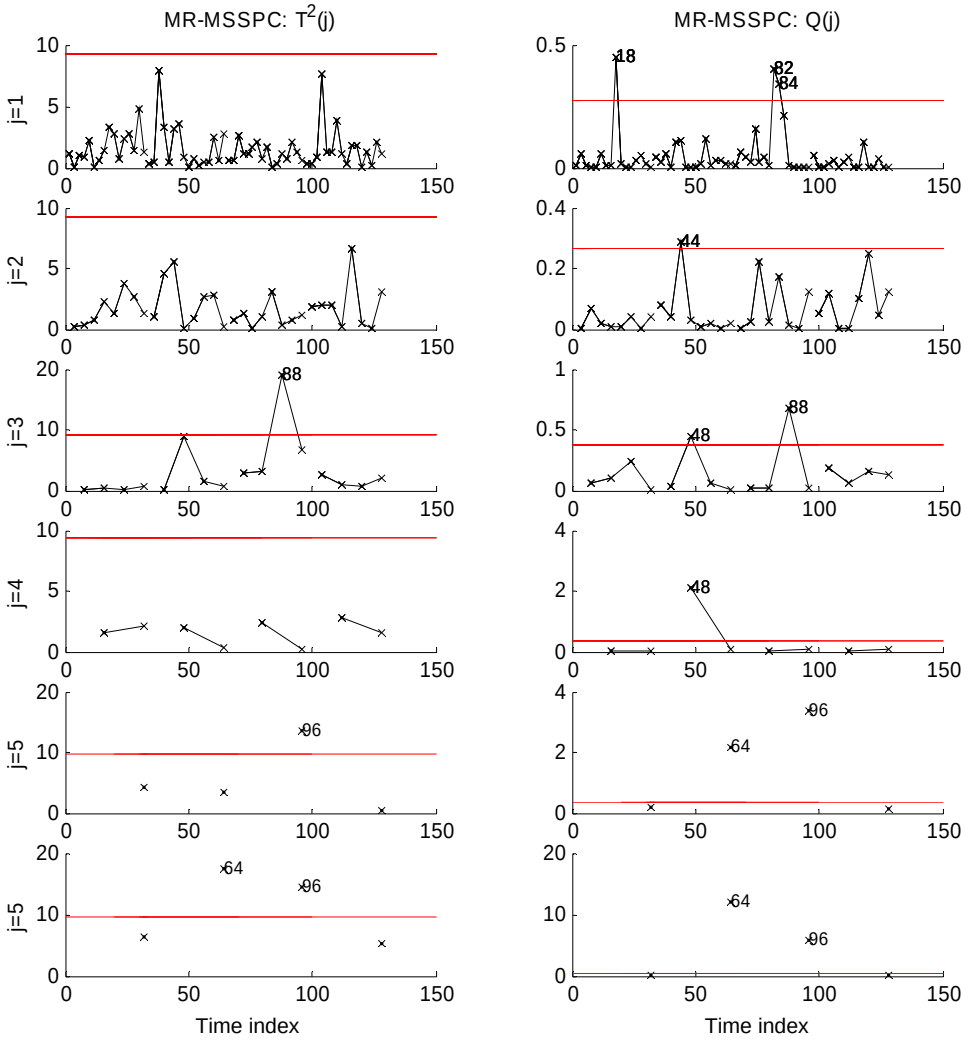


Figure 9.4. Plots of the T^2 and Q statistics for detail coefficients at each scale ($0 < j \le J_{dec}$) and for approximation coefficients at scale J_{dec}, with control limits set for a confidence level of 99%.

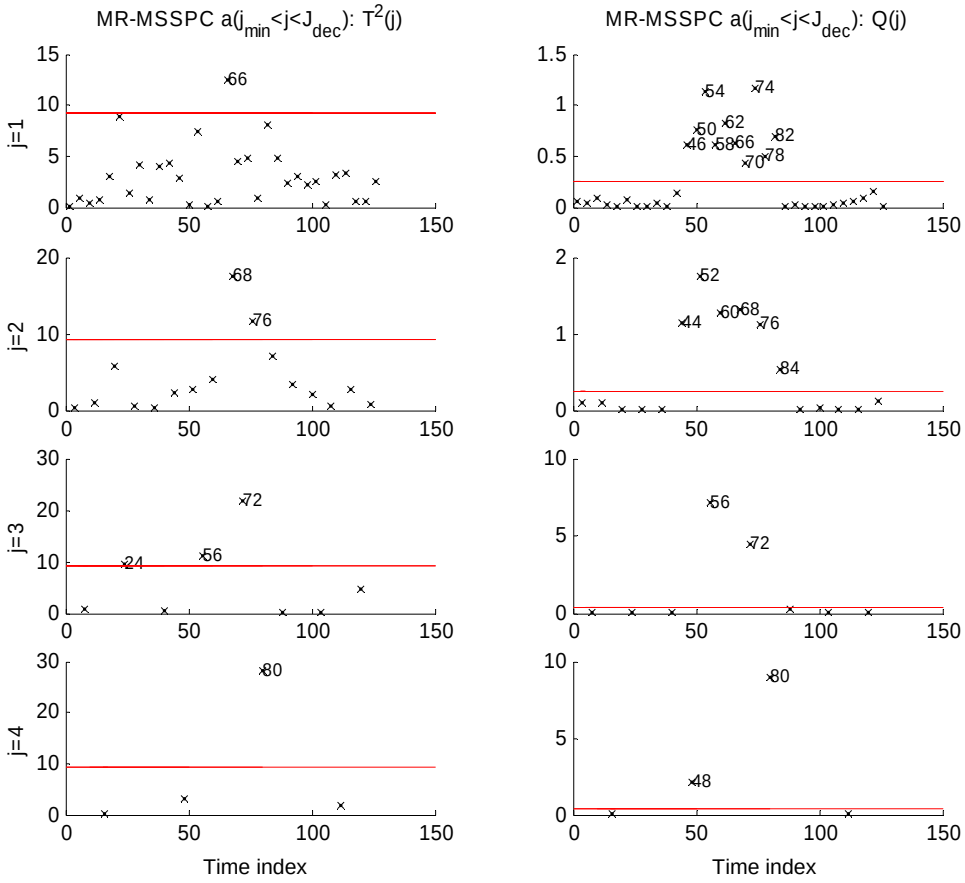


Figure 9.5. Plots of the T^2 and Q statistics for the approximation coefficients at scales $0 < j < J_{dec}$, with control limits set for a confidence level of 99%.

We now analyse the same situation, but using techniques designed to handle data at a single resolution that adopt the procedure for handling multiresolution data represented in Figure 9.2-b. Results regarding the T^2 and Q statistics for the MSSPC methodology with uniform discretization (Unif.-MSSPC) are presented Figure 9.6. Again, the number of the observation appears as a label when it is significant from the stand point of the chart statistic. One can see that control charts detect the shift quite promptly, but the definition of the region where the shift occurs is distorted, due to the way values for lower resolution variable are handled.

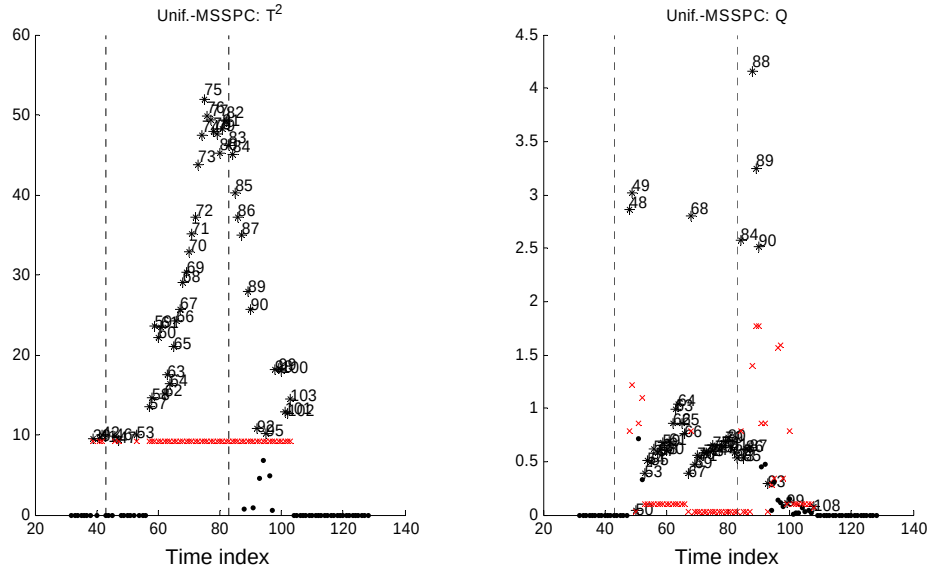


Figure 9.6. Results for MSSPC with uniform discretization: plots of the T^2 and Q statistics for reconstructed data. Control limits are set for a confidence level of 99% (represented by symbol x).

Figure 9.7 presents results regarding the use of PCA-SPC. We can see that in this case only the Q statistic detects significant abnormal activity going on during the duration of the shift, even though its detection rate is not as high as that exhibited by multiscale methods. This difference derives from the increased sensitivity of MSSPC methods, which have the ability of zooming into process behaviour at different scales (octave frequency bands), looking for changes in normal variability patterns.

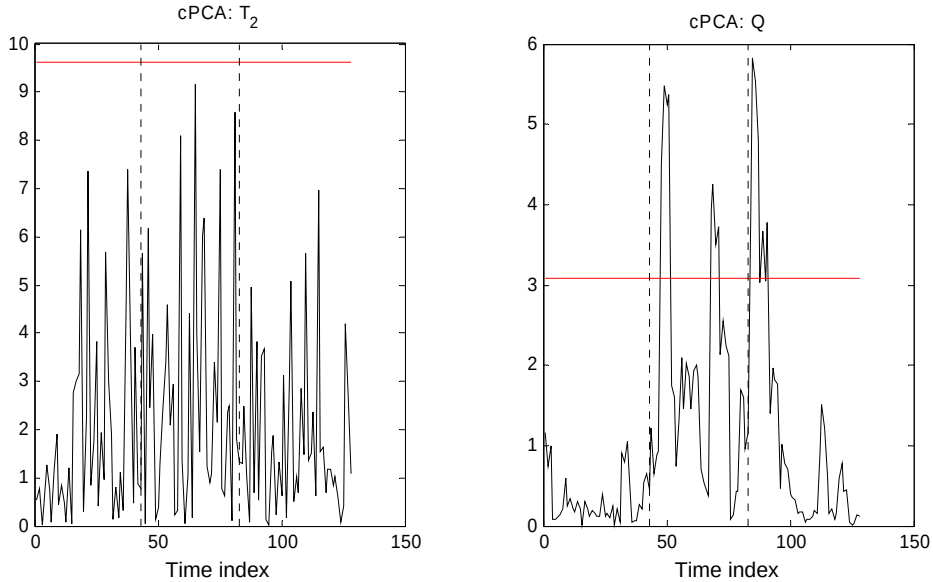


Figure 9.7. Results for cPCA-SPC: plots of the T^2 and Q statistics, with control limits set for a confidence level of 99% (cPCA stands for "classical" PCA, to distinguishing it from other related methods such as MLPCA; in this thesis, cPCA and PCA have the same meaning and are used interchangeably).

To sum up, the main feature of the MR-MSSPC methodology illustrated in this example is its ability to define more clearly the duration of the abnormalities when multiresolution data are present. It is quite sensitive in detecting its beginning, but, even more so, effective in the detection of its end, due to a consistent use of time supports regarding low resolution values achieved through the implementation of an orthogonal wavelet transformation over a variable dyadic length window. These features can be quite clearly seen in Figure 9.8, Figure 9.9 and Figure 9.10, where the time instants where significant events were signalled in the T^2 and Q control charts are presented as 1's, and the reminding non-significant or non-existent observation times, as 0's. These detection plots underline the delayed return to normal operation of the statistics in the Unif.-MSSPC method, and the better definition of the shift duration obtained with MR-MSSPC.

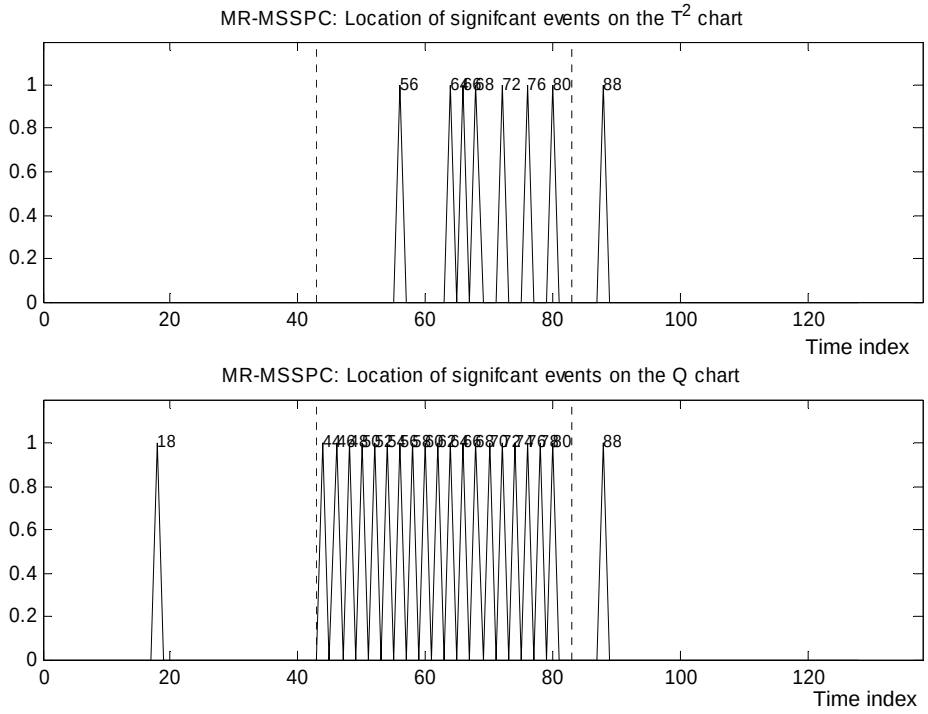


Figure 9.8. MR-MSSPC results: significant events detected in the charts for the T^2 and Q statistics (a significant event is signalled with "1").

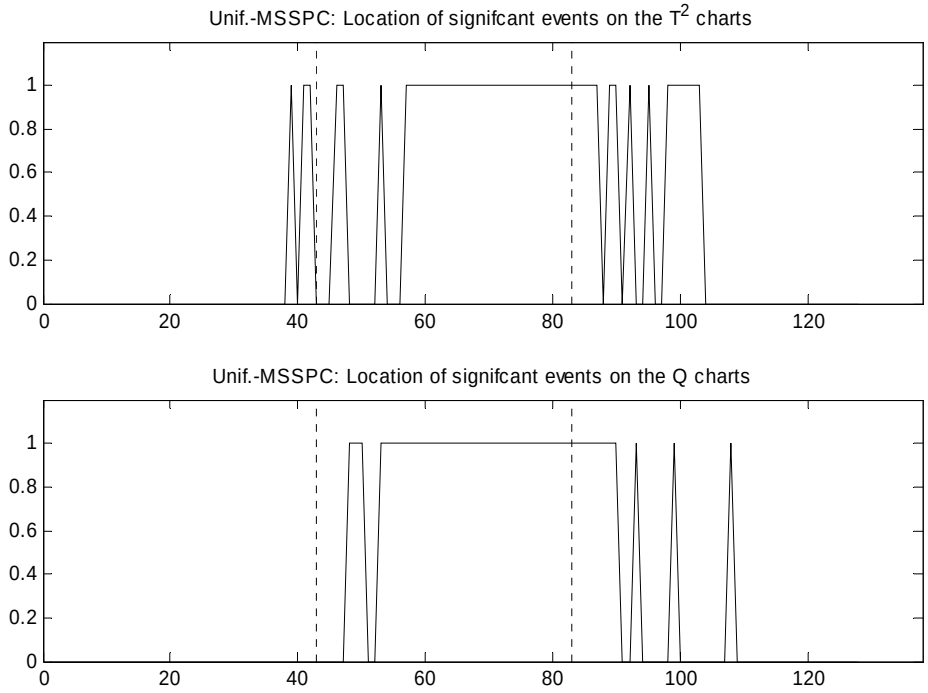


Figure 9.9. Unif.-MSSPC results: significant events detected in the charts for the T^2 and Q statistics (a significant event is signalled with "1"). Here, Unif.-MSSPC stands for the MSSPC methodology implemented with a uniform discretization scheme.

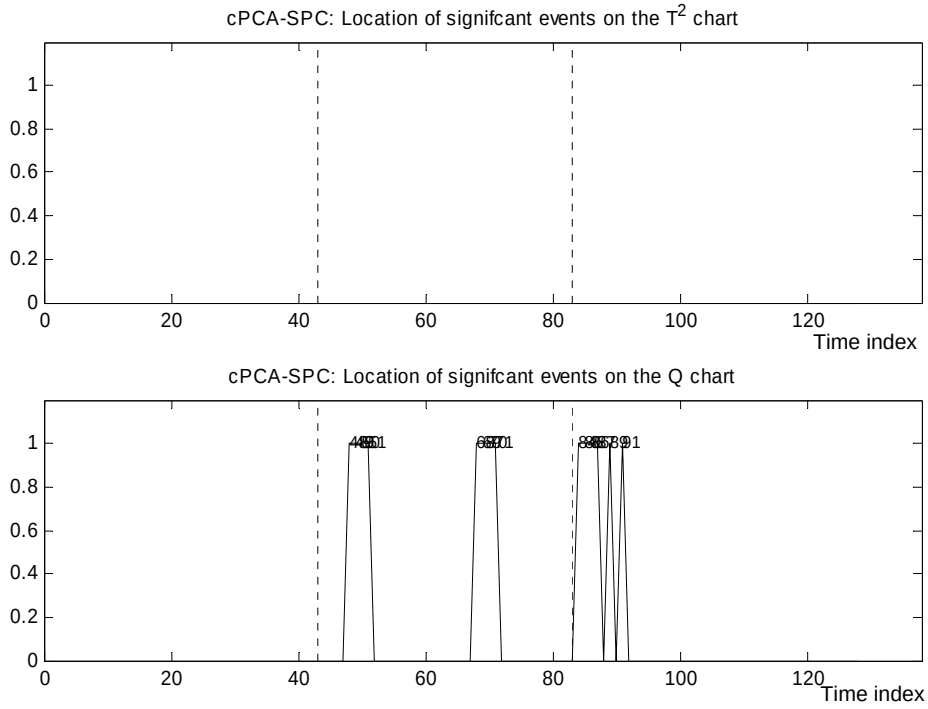


Figure 9.10. cPCA-SPC results: significant events detected in the charts for the T^2 and Q statistics (a significant event is signalled with "1").

9.4.2 Example 2: MR-MSSPC Extended Simulation Study

In order to consolidate the properties attributed to MR-MSSPC, which were illustrated in the previous example, an extended Monte Carlo simulation study was conducted, where several shifts were tested, together with different resolution levels associated to variable X_4 (or, stated equivalently, with averaging windows of different lengths for variable X_4). Resolution levels tested were $J_4 = \{2,3\}$, and the shift magnitudes analyzed were as follows: $shifts = \{0, 0.5, 1, 2, 3, 4\}$. For each resolution level, a reference data set composed of 2048 observations was created in order to estimate the models underlying each of the tested methodologies, after which a test set with 256 observations was generated with a shift introduced between observation number n_i and observation number n_f. To avoid biases due to shift location, n_i is randomly extracted from an uniform distribution, $n_i \sim U(40,50)$, while the duration of the shift was kept constant, corresponding to the next 40 observations ($n_f = n_i + 40$). The generation of the test set and shift location, was repeated 2000 times for each shift magnitude, and the results of the detection metrics saved for posterior calculation of average values.

The methods tested and compared are the following: MR-MSSPC, Dyadic-MSSPC (similar to MR-MSSPC with dyadic discretization strategy, but using data at a single resolution – the finest one), Unif.-MSSPC and PCA-SPC.

The detection metrics that will provide a ground for comparison are:

- Average run length (ARL), calculated considering the first occurrence of a significant event either in the T^2 or Q control charts;
- True Positive Rate (TPR), in this work corresponding to the fraction of significance events detected during the duration of the shift (between n_i and n_f), relatively to the maximum possible amount of detections that could be achieved with each methodology (i.e., if all the statistics' values computed during this time interval were significant);
- False Positive Rate (FPR), here defined as representing the fraction of false alarms detected right after the process returns to normality, in a range of time

with the same amplitude as the one used for calculating TPR (between $n_f + 1$ and $n_f + 41$), once again relatively to the maximum possible amount of detections that could be achieved with each technique.

The number of selected principal components was kept constant at 2, J_{dec} set equal to 5, and the wavelet transform used was the Haar transform. Significance levels adopted for each method were adjusted in order to obtain similar ARL(0) performances (average run length obtained under the absence of any shift), to enable for a fair comparison of the different methods involved.

Figure 9.11 compares ARL performances obtained for the various methods. The time delay associated with MR-MSSPC only becomes an issue for shifts of magnitude greater than 2, value after which it stabilizes at around 0.5, which seems acceptable for most applications. Thus, even though speed of detection was not a specific goal motivating the conception of our framework, it ends up performing well also in this regard.

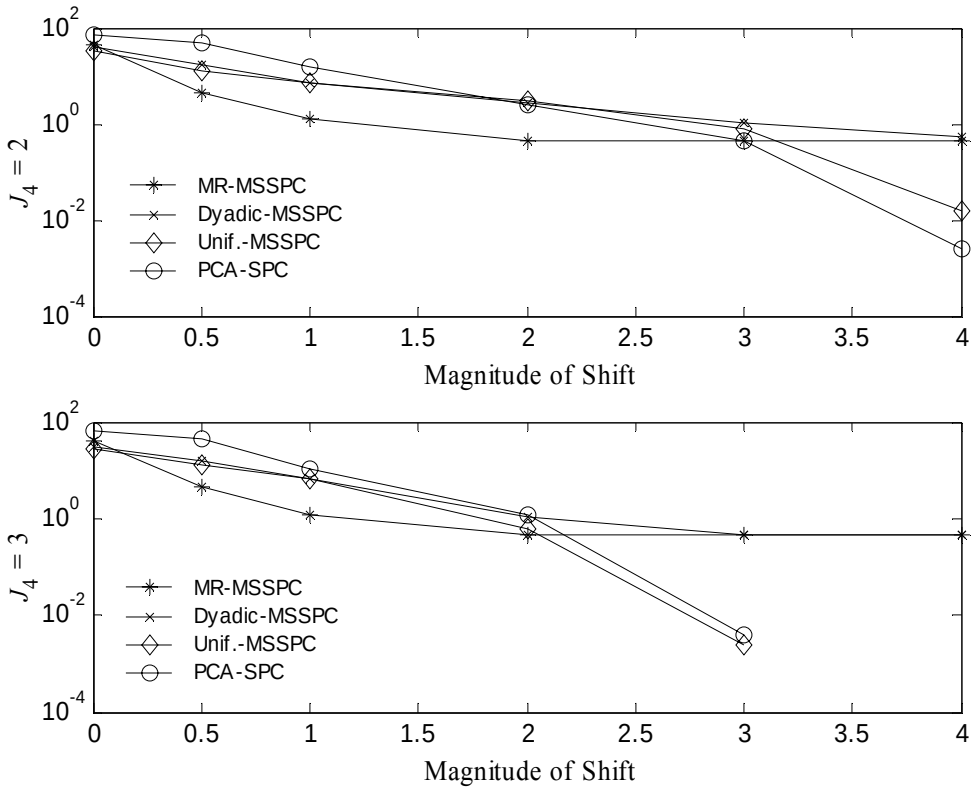


Figure 9.11. ARL results for the different methodologies, using shifts of different magnitude and two levels of resolution associated with variable X_4.

Results regarding TPR are shown in Figure 9.12, where we can see that MR-MSSPC performs better than its alternatives. It also does quite well when the process goes back to normal operation (Figure 9.13), with a low false alarm rate, which is only sometimes overtaken by PCA-SPC. However, in this situation, one must not forget that such a technique presents lower true positive detection metrics (TPR), as shown in Figure 9.12. Therefore, these results point towards an improved overall performance achieved by MR-MSPSC regarding the duration of the fault (higher TPR), quick detection of its beginning (low ARL) and effective delimitation of its end (low FPR).

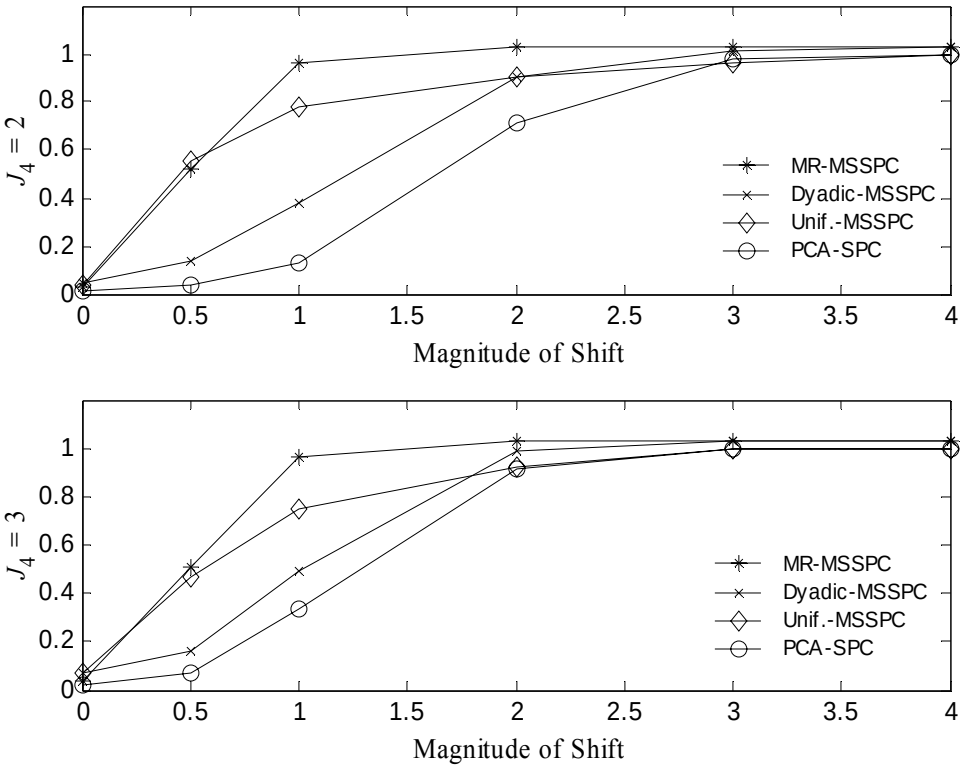


Figure 9.12. TPR results for the different methodologies, using shifts of different magnitude and two levels of resolution associated with variable X_4.

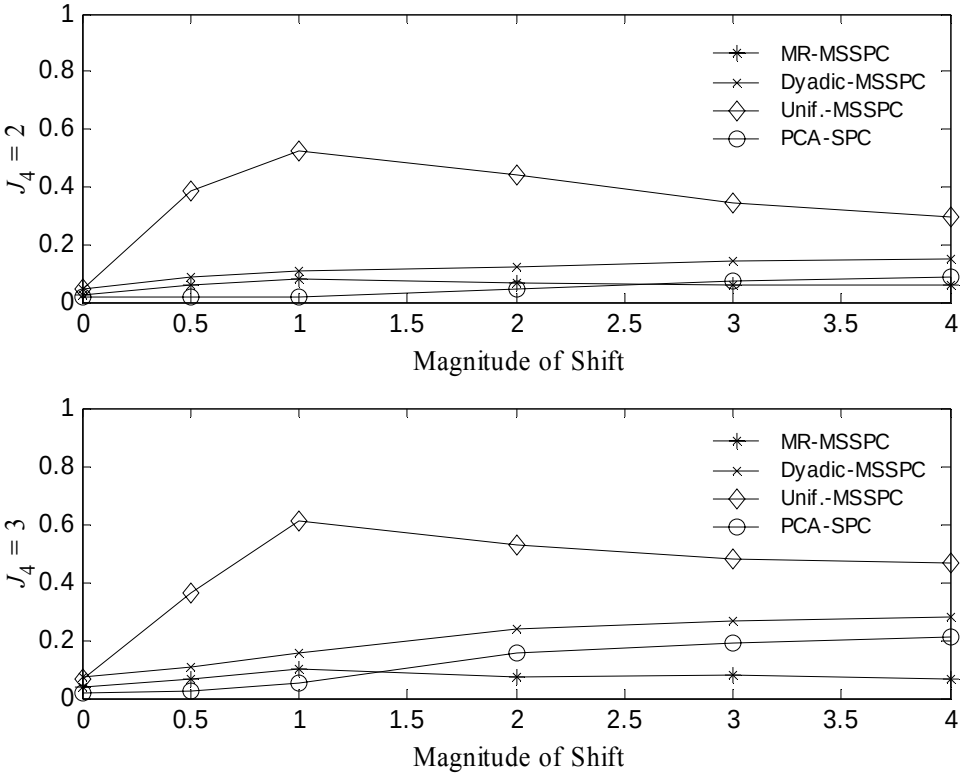


Figure 9.13. FPR results for the different methodologies, using shifts of different magnitude and two levels of resolution associated with variable X_4.

9.4.3 Example 3: MR-MSSPC for Multiresolution Data with Non-Dyadic Supports

When the values of a lower resolution variable represent the mean values over a non-dyadic time support, the attribution of its scale index is not straightforward. A simple way for handling this issue consists on implementing the steps presented in Table 9.2.

When the averaging supports have dyadic length, the above procedure provides the same values for X_i as the standard procedure. When they do not have such a property, it balances the contribution from each value within each sub-region of length 2^{J_i}, giving more weight to those that occupy a higher fraction of the interval.

Table 9.2. Selection of resolution index (J_i) when the averaging support for a lower resolution variable is not dyadic.

I. *Set J_i as the index for the next coarser scale, i.e. $J_i = \lceil \log_2(AS) \rceil$, where AS is the averaging time support and $\lceil \log_2(AS) \rceil$ stands for the smallest integer $n \geq \log_2(AS)$;*

II. *Project data onto scale J_i using the following weighted averaging procedure:*

a. FOR each window of dyadic length under analysis (length $2^{J_{max}(k)}$), divide it into sub-regions of length 2^{J_i} (K sub-regions);

1. FOR each sub-region ($l = 1:K$):

i. collect lower resolution values (x_j) and calculate the portion of their averaging support contained in the sub-region l under analysis (w_j);

ii. calculate the weighted average of the collected values: $X_i(l) = \sum_j w_j x_j \Big/ \sum_j w_j$;

END

END

To illustrate the application of this strategy, let us consider variable X_4 in model (9.1) to represent the average over a window of 5 successive values. Thus, according to step I in Table 9.2, $J_4 = \lceil \log_2(5) \rceil = 3$. The data set is also processed in order to be used with approaches based on single resolution data, by holding average values constant until a new mean value becomes available (Figure 9.2-b). The results obtained for MR-MSSPC, Unif.-MSSPC and PCA-MSSPC, when a shift of magnitude 1 is introduced between observation 43 and 83 (included), are presented in the plots from Figure 9.14 to Figure 9.19. Comparing Figure 9.14, Figure 9.15 and Figure 9.16 (or Figure 9.17, Figure 9.18 and Figure 9.19, that contain basically the same information, but where it is easier to identify the regions where significant events occur), we can verify an

improvement in the definition of the faulty region as well as in the detection of return to normality obtained through MR-MSSPC, even for this situation, where the averaging window does not have a dyadic support.

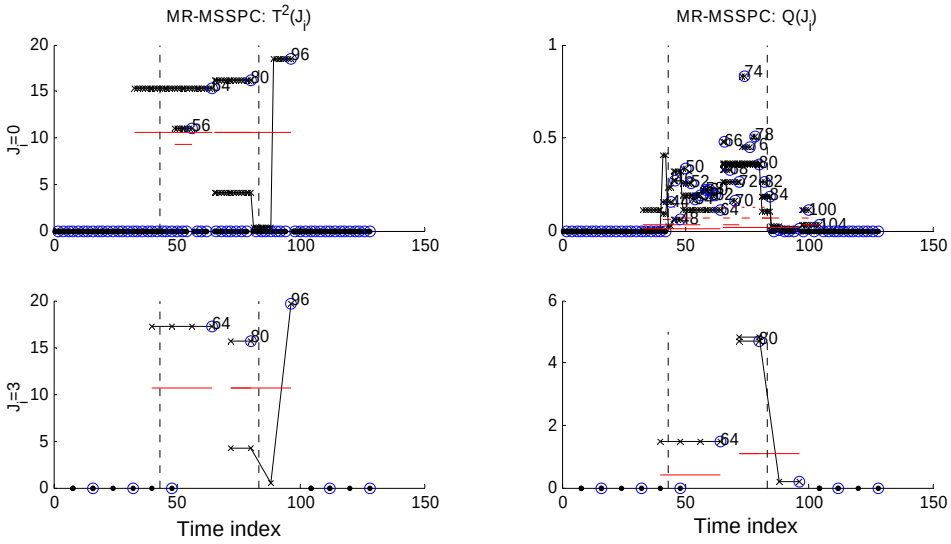


Figure 9.14. Plots of the T^2 and Q statistics at the two resolutions available in the data set, $J_i = \{0,3\}$, using data reconstructed from significant scales. Control limits are set for a confidence level of 99%.

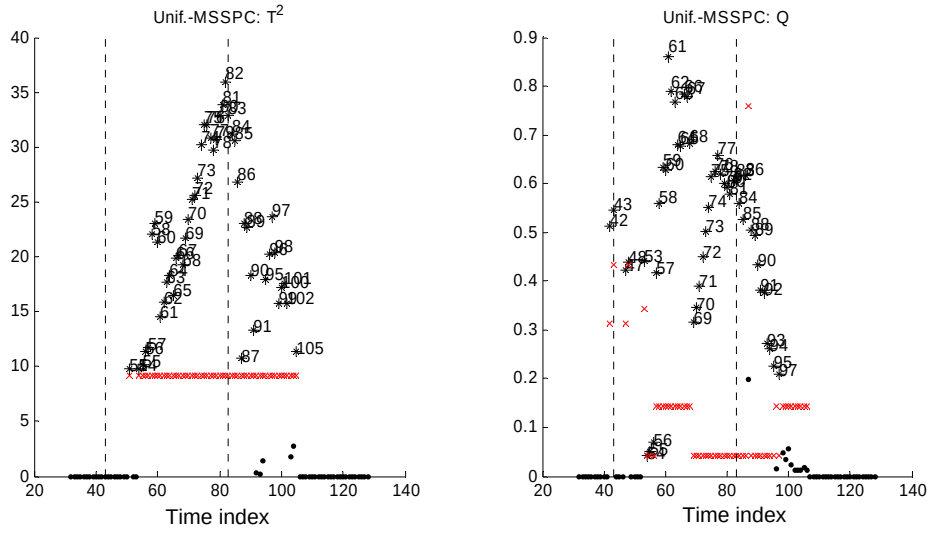


Figure 9.15. Results of MSSPC with uniform discretization: plots of the T^2 and Q statistics for the reconstructed data. Control limits are set for a confidence level of 99% (represented by symbol x).

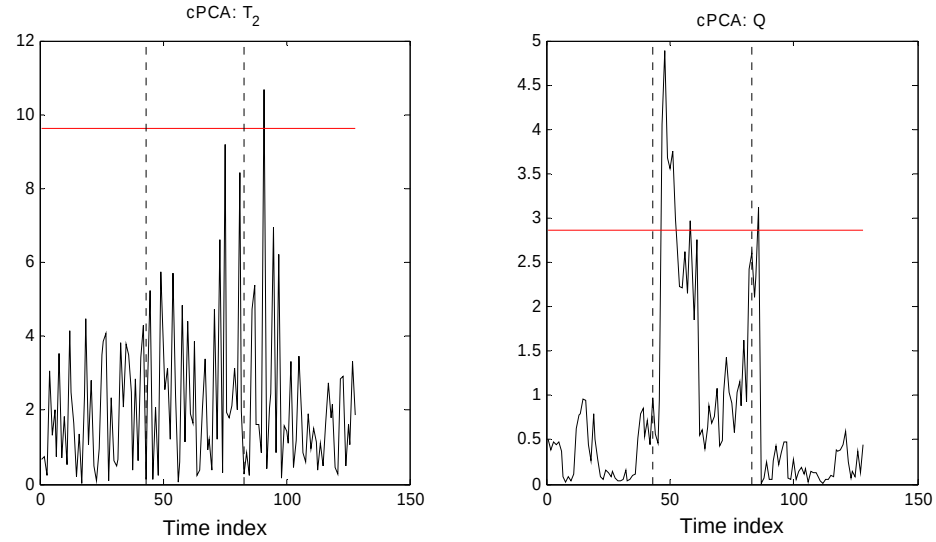


Figure 9.16. Results for PCA-SPC: plots of the T^2 and Q statistics. Control limits are set for a confidence level of 99%.

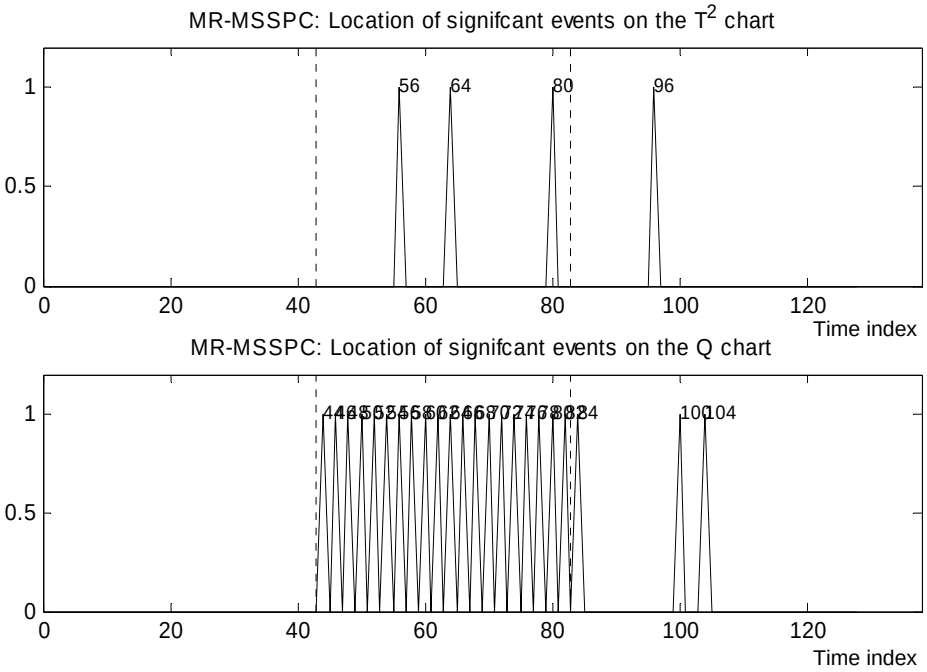


Figure 9.17. Results for MR-MSSPC: significant events detected in the charts for the T^2 and Q statistics (a significant event is signalled with "1").

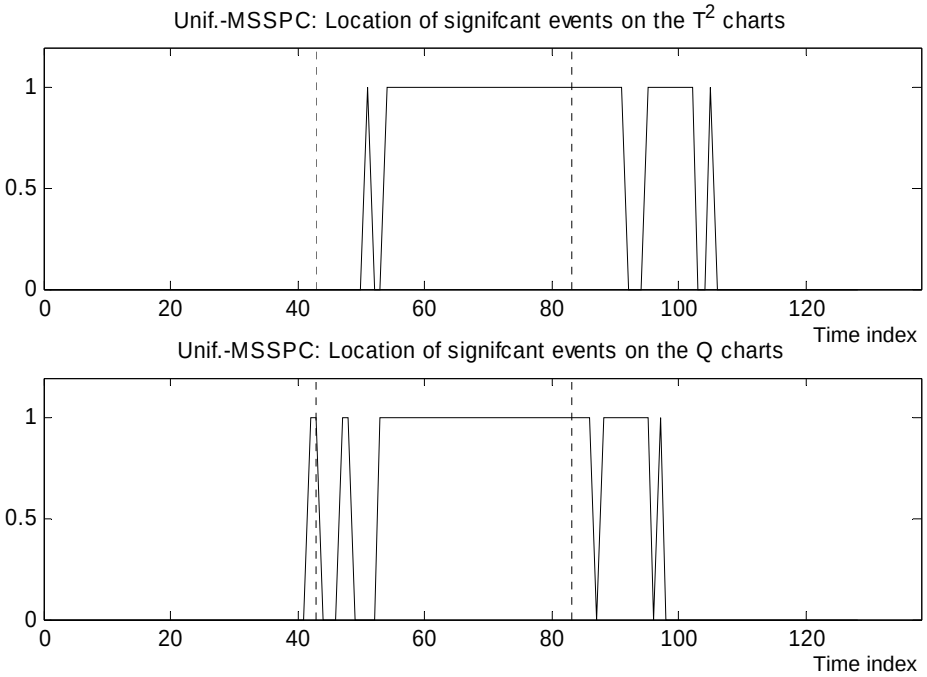


Figure 9.18. Results for Unif.-MSSPC: significant events detected in the charts for the T^2 and Q statistics (a significant event is signalled with "1").

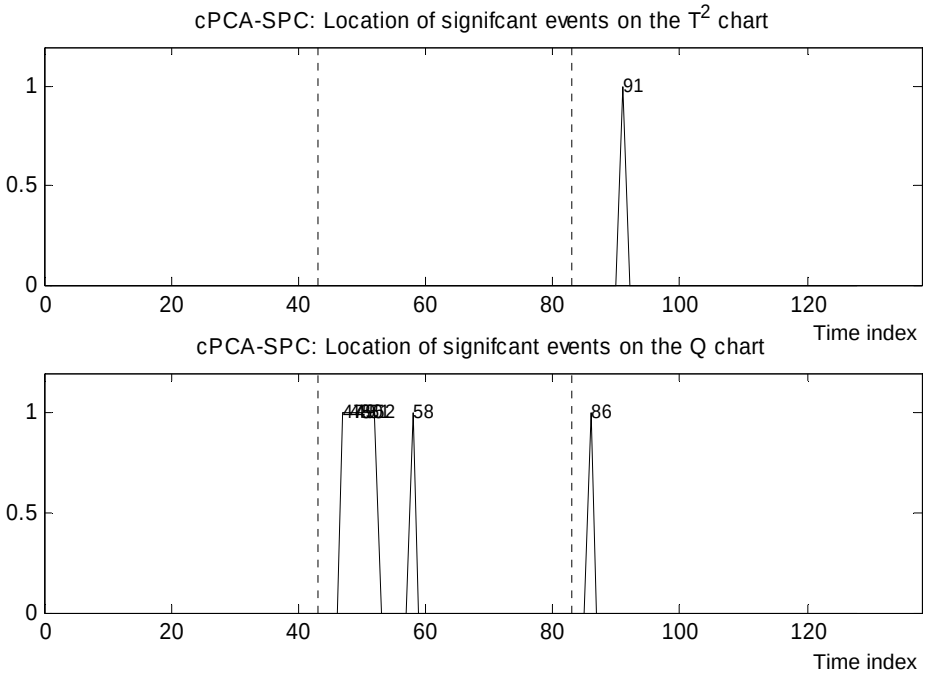


Figure 9.19. Results for PCA-SPC: significant events detected in the charts for the T^2 and Q statistics (a significant event is signalled with "1").

9.4.4 Example 4: MR-MSSPC Applied to a CSTR with Feedback Control

This final example aims to illustrate several interesting features that may arise in real world process monitoring implementations of MR-MSSPC, under the presence of time dynamics, non-linearity, besides variable collinearity. The system considered here consists of a simulated industrial non-isothermal CSTR, where an irreversible, exothermic first order reaction (A→B) takes place (Luyben, 1990). This reactor is equipped with a water jacket, that removes excess heat released, and two control loops (proportional action) that act upon two manipulated variables, i.e., outlet flow rate (F) and flow rate through the jacket (F_{cj}), in order to control the process variables volume, V, and reactor temperature, T, respectively. Figure 9.20 illustrates this system, whereas more details about its mathematical model, parameters and operating conditions can be found in Appendix E.

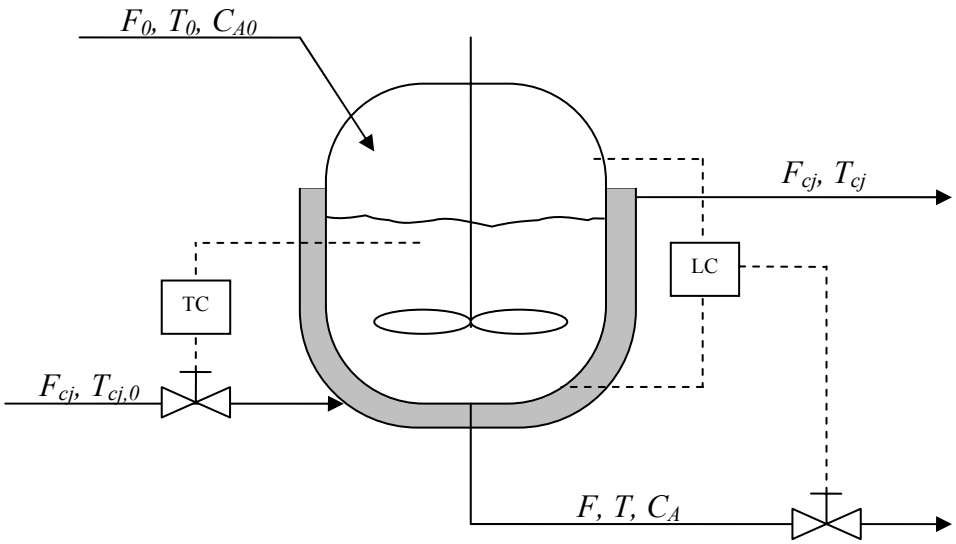


Figure 9.20. Schematic representation of CSTR with level and temperature control.

Normal operating conditions variability was generated by considering randomness associated with variables F_0, T_0, $T_{cj,0}$, and C_{A0}. For the first three variables, they were

assumed to be of the autoregressive type (first order, equation (9.2)), with parameters presented in Table 9.3.

$$X(k) = \phi \cdot X(k-1) + \varepsilon(k),\ \varepsilon(k) \sim N\left(0, \sigma_\varepsilon^2\right) \tag{9.2}$$

Table 9.3. Parameters of autoregressive models used for simulating normal operation regarding variables F_0, T_0, $T_{cj,0}$.

Variable	ϕ	σ_ε^2
F_0	0.5	0.750
T_0	0.95	0.878
$T_{cj,0}$	0.9	1.72

As for C_{A0}, it was assumed that the reagent is fed to the reactor from successive tanks, with $2\ m^3$ each, for which concentration shows little variation (approximately homogenous mixture of reagent in each tank), but that changes from tank to tank according to $C_{A0} \sim N\left(0.5, 0.1^2\right)$. All the measured variables are also subject to *i.i.d.* Gaussian noise.

A reference data set was generated representing 364 hours of normal operation, during which 10 variables, $\{X_i\}_{i=1:10} = \{V,\ T,\ T_{cj},\ C_A,\ F_0,\ C_{A0},\ T_0,\ T_{cj,0},\ F,\ F_{cj}\}$, were collected every 10 s, and analyzed in order to set monitoring parameters, estimate models at each scale and gain insight regarding multiscale features (illustrations of the time series plots associated with each variable are also presented in Appendix E).

In this preliminary analysis stage, a decomposition depth of $J_{dec} = 12$ was chosen, which is high enough to characterize all the phenomena going on for this process. Figure 9.21 presents the eigenvalues profile for covariance matrices at each scale, and Figure 9.22 shows the cumulative percentage of explained variance for each new component considered in the PCA model developed at each scale (all variables were previously "autoscaled", i.e., centred at zero and scaled to unit variance). These plots clearly illustrate that the dimension of the relevant PCA subspaces for process monitoring purposes, in dynamic non-linear systems, is in general a function of scale,

according to the power spectra of the variables involved and the correlation they present in the frequency bands corresponding to different scales. This is in contrast with what is expected to happen with stationary uncorrelated processes, such as (9.1), where the covariance structure should be the same at all scales, up to a constant multiplier term related to the square gain function of the wavelet filter, since the cross spectral density function of the variables (Whitcher, 1998) is, in this case, constant throughout the whole frequency spectrum. Using the information conveyed by such plots, we can also choose the number of principal components to be adopted for PCA models of the wavelet coefficients at each scale, as well as get important clues regarding the decomposition depth that should be used in order to capture the system's main dynamical features. In this particular case the decomposition depth was set as $J_{dec} = 9$, because above this scale the behaviour of the correlation structure does not seem to change significantly. This means that all relevant dynamic features of the system are expressed at lower scales.

The absolute values of the loading vectors, for the selected principal components at each scale, are presented in Figure 9.23 (shadowed plots), where we can deepen the analysis of the correlation structure at different scales, looking for the main active relationships in each frequency band, and distinguish which variables are more significantly involved. This last point can be conducted more effectively by looking at the percentage of explained variance for each variable in the PCA model developed at each scale (Figure 9.24). From these two figures we can see that, although there is some overlapping due to the interception between frequency bands characteristic of some variables, in scales 1-3 the variables involved are mainly those with fast dynamics (notably flow rates), whereas in the intermediate scales (3-8) we get those variables regarding disturbances with slower dynamics (temperatures), as well as the attenuated effect of flow rate "filtered" by reactor capacity (volume). Finally, in scales 8-12, the slow mode variables (C_{A0}, the majority of the system outputs, and control variables that react based upon the measured values of the outputs) become relevant.

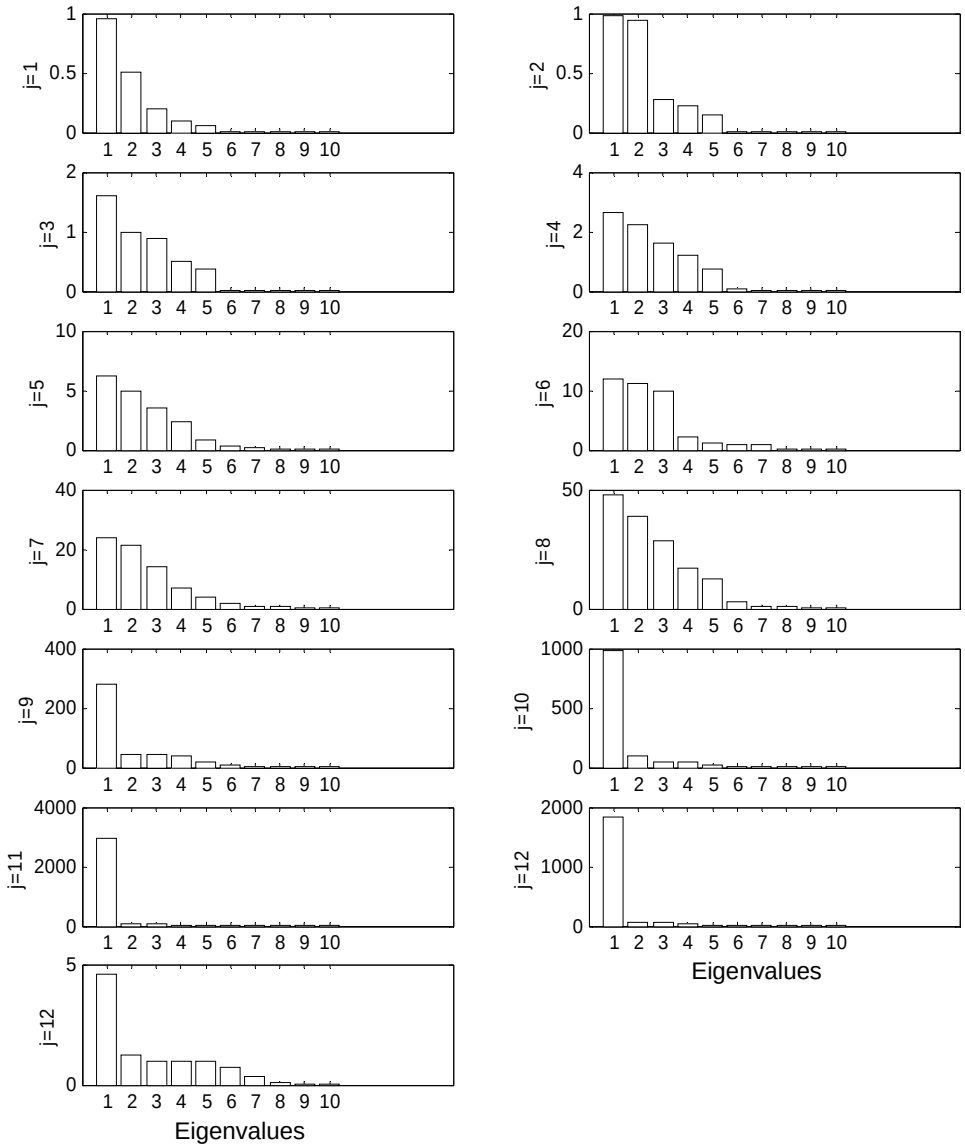


Figure 9.21. Eigenvalue plots for the covariance matrices regarding variables' wavelet detail coefficients at each scale ($j = 1:12$) and for the wavelet approximation coefficients at the coarsest scale ($j = 12$, last plot at the bottom).

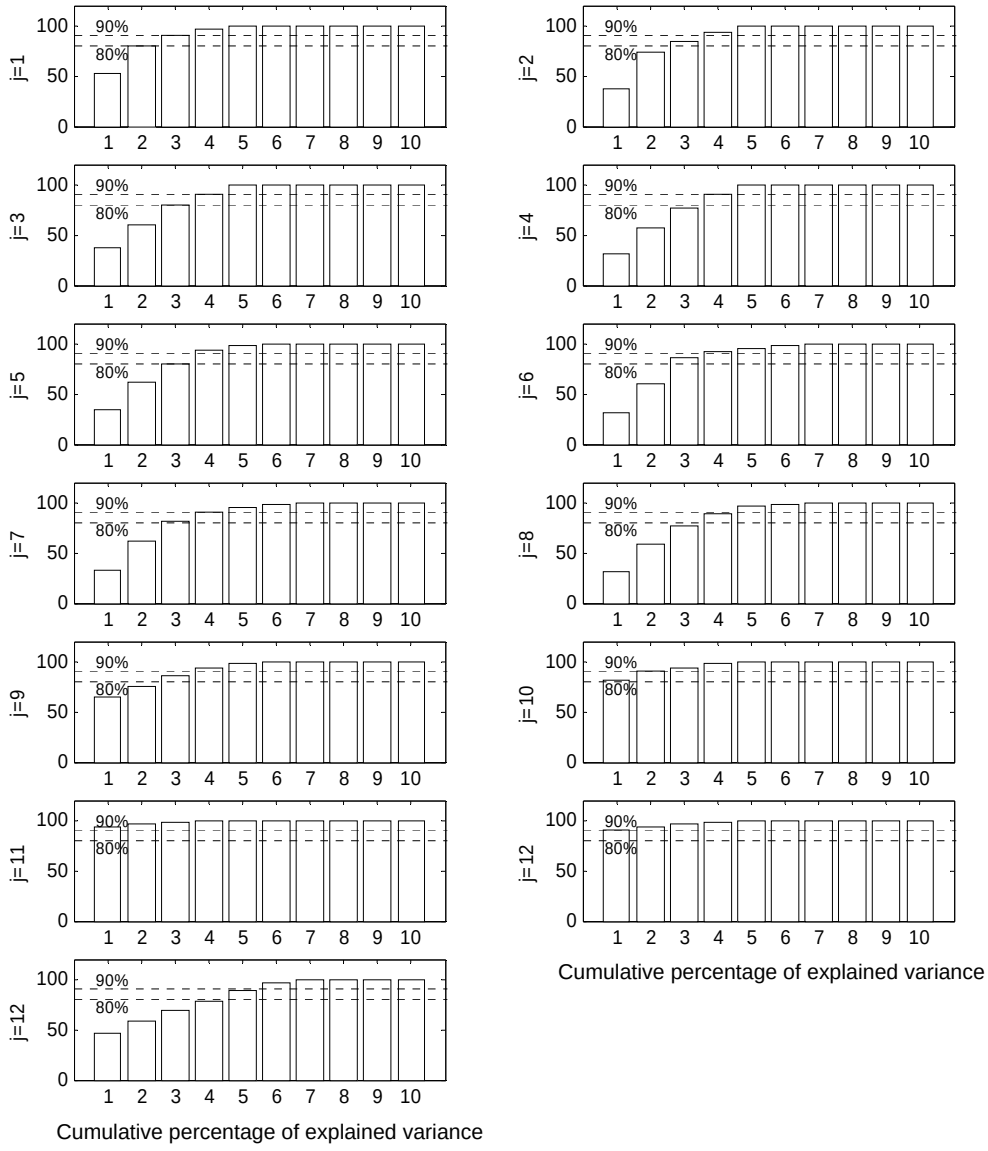


Figure 9.22. Plots of the cumulative percentage of explained variance for each new component considered in a PCA model developed at each scale, for the detail coefficients ($j = 1:12$) and approximation coefficients at the coarsest scale ($j = 12$, last plot at the bottom).

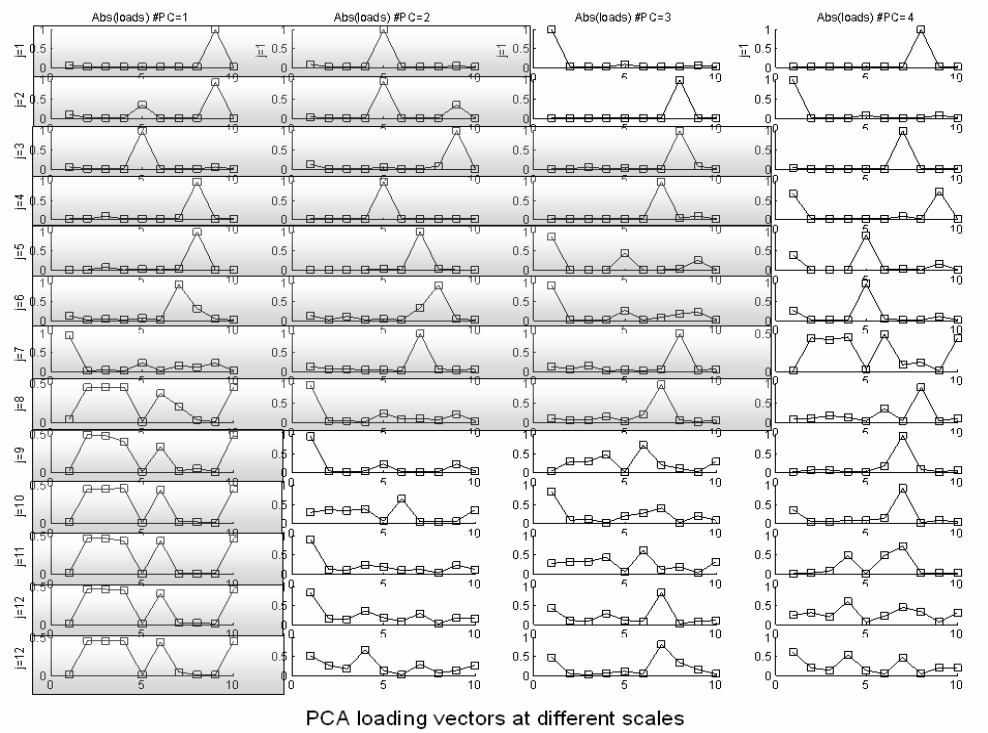


Figure 9.23. Absolute values of the coefficients in the loading vectors associated with the principal components selected at each scale (shadowed graphs).

After having estimated the MR-MSSPC monitoring parameters, a test data set containing about 45 hours of operation data was generated, with a bias of 6 K introduced in T_0 ($\Delta/\sigma \approx 2$), between times 22h:46m and 34h:09m. The monitoring results obtained for the MR-MSSPC Q statistic are presented in Figure 9.25, showing that the proposed method successfully detected such a shift. As Figure 9.24 points out, variable number 4, C_{A0}, only becomes relevant at coarser scales. This means that we can use a lower resolution to represent its behaviour along time, without loosing much detail but introducing a time delay in the decision-making process associated with such a variable. Figure 9.26 illustrates what happens when we set the resolution of C_{A0} at $J_4 = 5$, and conduct MR-MSSPC over the same test data set. The detection results do not change significantly, but the location of the faults becomes even more evident in the representation at $J_i = 5$.

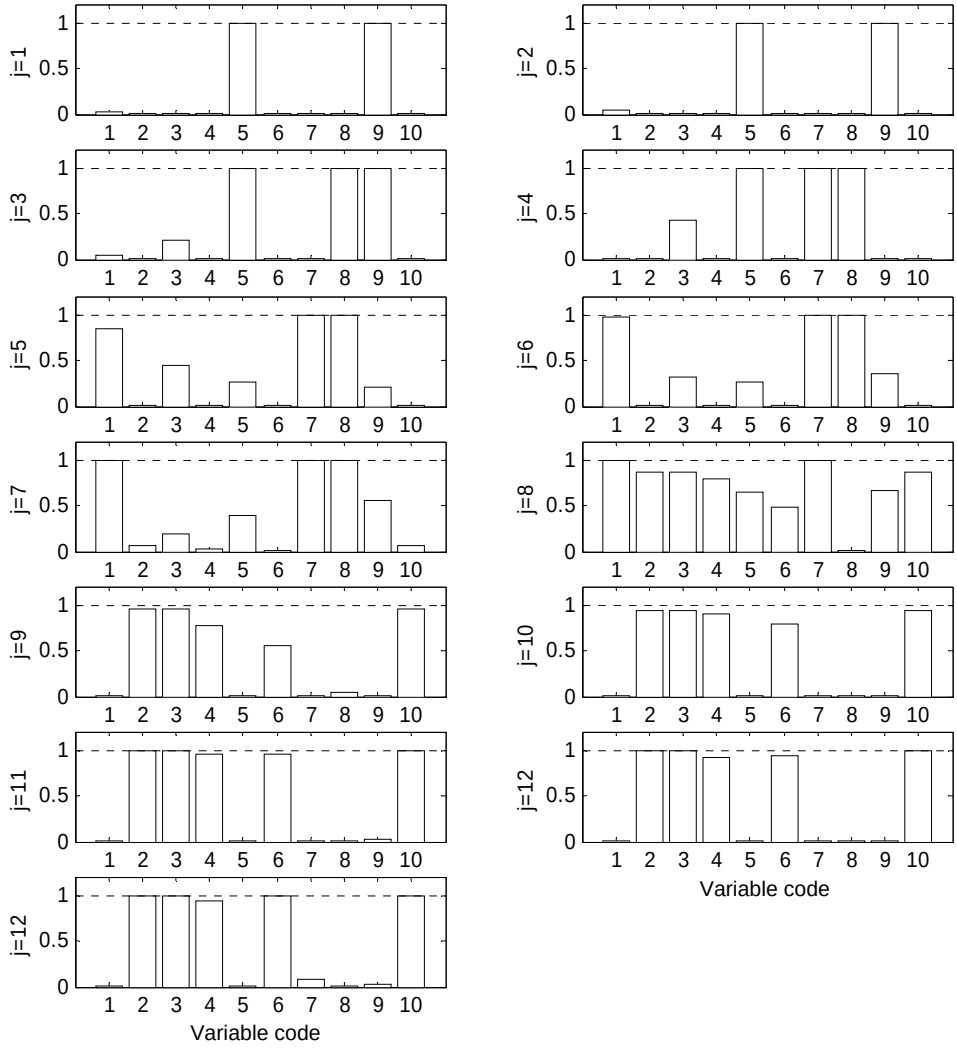


Figure 9.24. Percentage of explained variance for each variable in the PCA model developed at each scale.

In Figure 9.26 we are not only handling existing multiresolution data, but are actually creating a multiresolution data structure, after analysing the multiscale characteristics of the system operating under normal conditions. The coarser scale selected represents a trade-off between the adequate scale to express a certain variable and the time delay involved in the computation of its mean values, which may introduce a detection delay for the special case where a fault is only present in this particular variable.

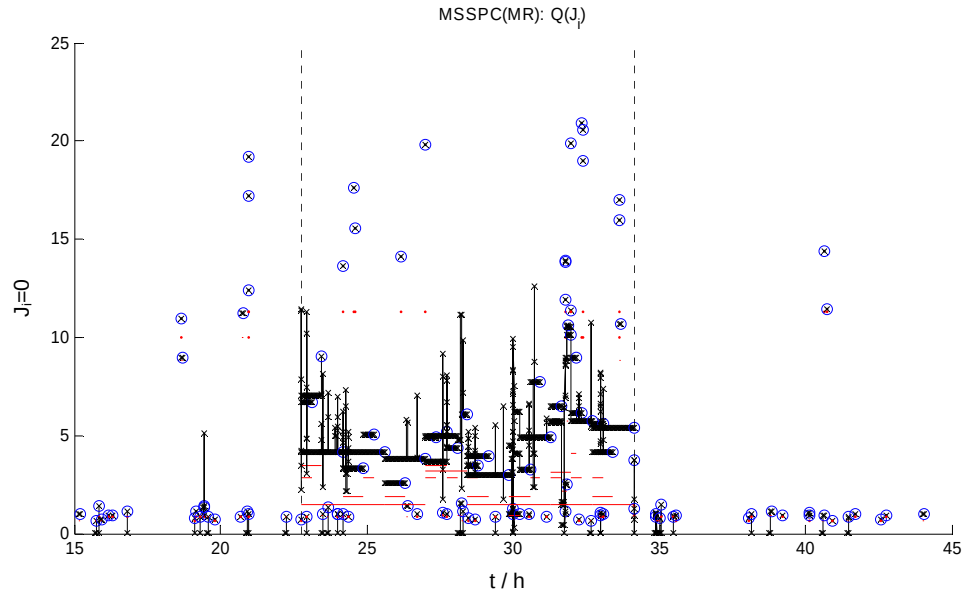


Figure 9.25. Plot of the Q statistic for MR-MSSPC applied over the test data set, with all variables available at the finest scale $J_i = 0$ $(i = 1:100)$ (Control limits defined for a 99% confidence level).

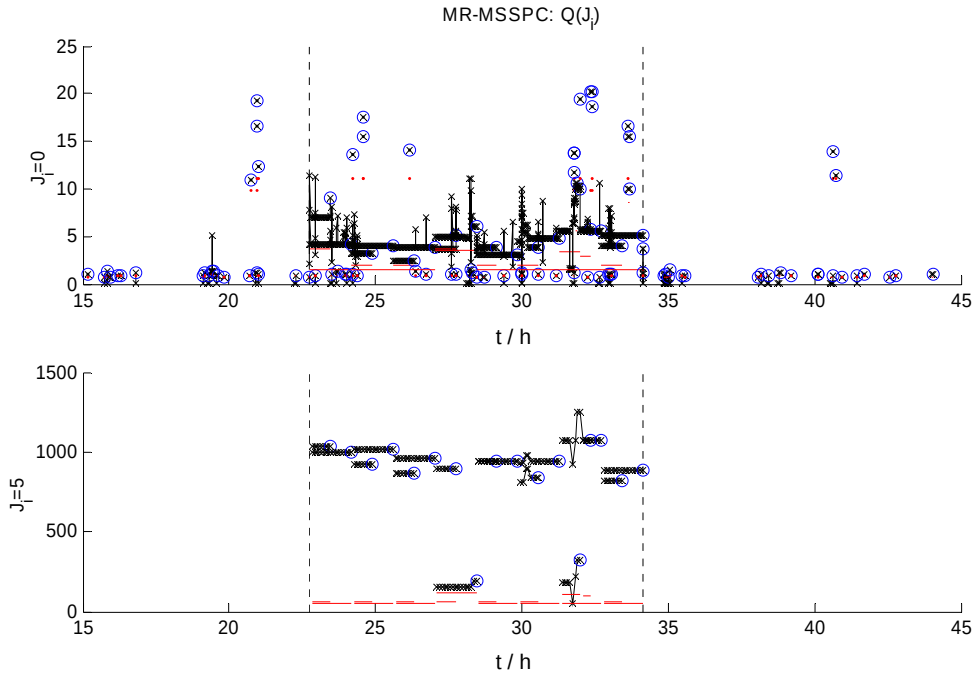


Figure 9.26. Plot of the Q statistic for MR-MSSPC applied over the test data set, with all variables available at the finest scale, except for C_{A0}, which is now only available at $J_4 = 5$ (control limits defined for a 99% confidence level).

A delay may also be present when the above situation occurs for a (conventional) multiresolution data structure. To illustrate better this issue, let us consider the case where a low resolution variable, involving averages over large data windows, suffers a fault that only manifests itself in this particular variable and begins almost at the end of the last averaging period. Then, this fault may pass unnoticed during the current monitoring stage (as it represents very little of the information averaged), which implies that it will only be detected after 2^{J_i+1} time steps (if it is active during a part of this period). Even though detection speed is not one of the main features advocated for MR-MSSPC (although it performs quite well in this regard, as shown in the previous examples), this point can be improved by adopting an hybrid approach, between the proposed approach for handling multiresolution data and the conventional one, that consists on artificially augmenting the sampling rate of this variable, by assuming the last average values to be constant in the mean time (just as in Figure 9.2-b), forcing the T^2 and Q statistics calculations to integrate the low resolution variable more often. This will sacrifice a bit the claimed definition ability regarding duration of the fault, in order to improve the promptness of detection for unusual events, that only affect a given isolated low resolution variable, but should only be implemented if this type of fault is really likely to occur.

9.5 Conclusions

In this chapter, a methodology was presented for conducting MSSPC that adequately integrates data with different resolutions (multiresolution data). Such an approach was then tested under four different scenarios in order to illustrate its main features. The first three examples underline consistent use of the time support regarding lower resolution variables, enabling for a clearer definition of the regions where significant events occur and a more sensitive response when the process is brought back to normal operation. They also show that, as long as the fault does not happen exclusively in the lower resolution variables, no significant time delay is introduced by the proposed methodology. A final example brings to the discussion both interesting and important issues regarding practical applications that involve dynamic systems with non-linear behaviour, such as the interpretation of their multiscale covariance structure and the selection of monitoring parameters.

Part V

Conclusions and Future Work

We are at the very beginning of time for the human race. It is not unreasonable that we grapple with problems. But there are tens of thousands of years in the future. Our responsibility is to do what we can, learn what we can, improve the solutions, and pass them on.

Richard P. Feynman (1918-1988), American theoretical physicist and educator.

Chapter 10. Conclusions

In this thesis, we have addressed the general problem of exploring information available at a given scale, or set of scales, in order to devise improved strategies for performing data analysis. Special attention was paid to certain relevant aspects in practical applications, like handling sparse data sets and the integration of uncertainty information in data analysis. In fact, sparsity (due to missing data or different acquisition rates) often hinders conventional analysis of industrial data sets to proceed smoothly or, at least, prevents the full exploitation of all information potentially contained in such data sets. On the other hand, with the development of measurement technology and metrology, we often really know more about data than just their raw values, since the associated uncertainty is (or, if that is not the case, "should be") also available. This means that methodologies based strictly on raw values can be now improved, through the integration of uncertainty information in their formulations, because this piece of information is becoming increasingly available. Another complicating feature that calls for attention in the analysis of industrial data sets is the presence of data at multiple resolutions (multiresolution data), whose averaging supports should be adequately incorporated in data analysis.

These difficulties and features were considered in the methodologies that were presented in this thesis, regarding the development of an industrial data-driven multiscale analysis framework. In the following paragraphs, a concluding summary reviews the new contributions proposed here, with the main *conceptual* outputs being presented in Table 10.1. On the other hand, Table 10.2 provides a list of *application-oriented* contributions and resumes topics where the emphasis lies either in application scenarios or in the type of tools used to address them.

Table 10.1. Summary of the thesis' main new *conceptual* contributions, along with references where they are, partially or thoroughly, treated (when applicable).

Contribution	*Section*	*Short Description*	*Reference*
MRD: Method 1	4.1.1	A generalized MRD framework	(Reis & Saraiva, 2005b)
MRD: Method 2	4.1.2	A generalized MRD framework	(Reis & Saraiva, 2005b)
MRD: Method 3	4.1.3	A generalized MRD framework	(Reis & Saraiva, 2005b)
Theorem 4.1	4.1.3	Covariance of wavelet-transformed noise	
Method MLMLS	5.1.1	Uncertainty-based linear regression methods	(Reis & Saraiva, 2005c)
Method rMLS	5.1.2	Uncertainty-based linear regression methods	(Reis & Saraiva, 2004b, 2005c)
Method rMLMLS	5.1.2	Uncertainty-based linear regression methods	(Reis & Saraiva, 2005c)
Method MLPCR2	5.1.3	Uncertainty-based linear regression methods	(Reis & Saraiva, 2005c)
Method uPLS1	5.1.4	Uncertainty-based linear regression methods	(Reis & Saraiva, 2004b, 2005c)
Method uPLS2	5.1.4	Uncertainty-based linear regression methods	(Reis & Saraiva, 2004b, 2005c)
Method uPLS3	5.1.4	Uncertainty-based linear regression methods	(Reis & Saraiva, 2005c)
Method uPLS4	5.1.4	Uncertainty-based linear regression methods	(Reis & Saraiva, 2005c)
Method uPLS5	5.1.4	Uncertainty-based linear regression methods	(Reis & Saraiva, 2005c)
uNNR	11.4	Uncertainty-based non-parametric regression	(Reis & Saraiva, 2004a)
Formulation I	6.1	Process optimization using data uncertainty	(Reis & Saraiva, 2005c)
Formulation II	6.1	Process optimization using data uncertainty	(Reis & Saraiva, 2005c)
HLV-MSPC	7.1-7.3	MSPC using data uncertainty	(Reis & Saraiva, 2003, 2005a)
MS monit. profiles	8.1	Multiscale monitoring of profiles	(Reis & Saraiva, 2005d, 2005e)
MR-MSSPC	9.3	MSSPC with multiresolution data	

Table 10.2. Summary of the thesis' main *application-oriented* contributions.

Contribution	*Section*	*Reference*
Uncertainty-based de-noising	4.3	(Reis & Saraiva, 2005b)
Implementation of scale selection methodologies	4.4	
Process optimization using uncertainty information	6.2	(Reis & Saraiva, 2005c)
Retrospective analysis of process data using HLV-MSPC	7.4.2	(Reis & Saraiva, 2005a)
Supervised classification models of paper surface quality	8.2.2	(Reis & Saraiva, 2005f, 2005g)
Time series modelling of roughness phenomena	8.2.3	
Multiscale monitoring of paper surface profiles	8.2.4	(Reis & Saraiva, 2005e)
MR-MSSPC applied to a CSTR under feedback control	9.4.4	

Several multiresolution decomposition (MRD) frameworks, that play an essential role when focusing data analysis at a particular scale, or set of scales, were developed, with the ability of incorporating uncertainty information and handling missing data

structures, features that are absent from classical MRD approaches based on wavelets (*Methods 1*, *2* and *3*, Section 4.1). Guidelines were also provided regarding their application in practical contexts (Section 4.2). Besides extending the wavelet-based multiresolution decomposition to contexts where it could not be applied otherwise (at least without some serious data pre-processing efforts), such methods also provide new tools for addressing other classes of problems in data analysis, such as the one of selecting a proper scale (Section 4.4). They also provide, in particular, data and associated uncertainty tables at a single-scale, that can be adequately handled by the methods described in Chapters 5 and 7.

The integration of uncertainty information in several data analysis tasks was explored in Part IV-A of this thesis. Several linear regression models were compared, some of them with the ability of incorporating uncertainty information in their formulations, including some new proposed methodologies,[41] and their performances compared using extended Monte Carlo simulations (Chapter 5). Under the conditions covered in the study, method MLPCR2 presented the best overall predictive performance and, in general, those methods based on MLMLS tend to present improvements over their counterparts based on MLS.

The use of measurement and actuation uncertainties in process optimization problems was also explored in Chapter 6, and several possible optimization formulations were analysed, differing on the levels of incorporation of uncertainty information. The analysis of results points out the relevance of not neglecting measurement and manipulation uncertainties when addressing both on-line and off-line process optimization.

Another task where uncertainty information was integrated, and turned out to provide an effective and coherent way to approach missing data, was in multivariate statistical process control (MSPC). In this context, a suitable statistical model was defined in Chapter 7 (HLV) and statistics analogous to T^2 and Q derived, that allow for monitoring both within model variability as well as variability around the estimated

[41] Namely: MLMLS, rMLS, rMLMLS, MLPCR2, uPLS1, uPLS2, uPLS3, uPLS4, uPLS5.

model. Results obtained point out in the direction of using such an approach, when noise has low SNR and uncertainties vary across time.

Two multiscale monitoring approaches were presented in Part IV-B. The first methodology, presented in Chapter 8, regards the monitoring of profiles, and is built around a wavelet-based multiscale decomposition framework that essentially conducts a multiscale filtering of the raw profile, effectively separating the relevant phenomena under analysis, located at different scales, allowing also for the incorporation of available engineering knowledge and information derived from the analysis of the distributions of different related quantities through the scales. The results presented for a specific case study, which deals with monitoring of paper surface using profilometry, allow us to conclude in favour of the adequacy of adopting the proposed approach for monitoring simultaneously the two relevant surface phenomena under study (roughness and waviness). Multiscale characteristics of paper surface were also carefully analyzed using specialized plots and time series theory. The availability of parameters provided by the measurement device was also explored for predicting classification of paper surface quality, through adequate classification models that explain assessments made by a panel of experts.

The second methodology, addressed in Chapter 9, provides a way for conducting MSSPC by adequately integrating data at different resolutions (multiresolution data). The proposed approach was tested under different scenarios, and we verified that the consistent use of time supports regarding lower resolution variables made by MR-MSSPC led to a clearer definition of the regions where significant events occur and a more sensitive response when the process is brought back to normal operation, when compared to the approaches based on single resolution data at the finest scale.

Chapter 11. Future Work

In this chapter, several research topics are addressed, representing examples of interesting areas for future research within the scope of the matters and results presented in this thesis. They are organized in different sections, according to the following fields:

- Multiscale black-box modelling and identification;
- Multiscale monitoring;
- Hierarchical modelling of multiresolution networks;
- Further developments on uncertainty-based methodologies.

11.1 Multiscale Black-Box Modelling and Identification

A wide variety of model structures is available for modeling the dynamical and stochastic behavior of systems, using data collected from industrial plants. Some well known examples are state-space, time series (ARMAX, Box-Jenkins), latent variables (PLS) and neural networks models. These model structures are however inherently single-scale, as they are used to address the modeling task from the stand point of developing an adequate description of reality regarding what happens strictly at the time scale corresponding to the adopted sampling rate. Information contained at other scales is not explicitly considered in these formulations, and therefore is frequently overlooked in methodologies based on model structures identified within the scope of such classes (e.g. process optimization, optimal estimation, fault detection and diagnosis). Therefore, the development of model structures that present the ability of explicitly integrating the

concept of "scale" in their core can potentially lead to better descriptions for systems that present scale-dependent dynamic behavior.

In this context, the development of new model structures, that *extend* the classic arsenal for discrete-time models to situations where the aforementioned limitations are relevant, provides interesting research challenges. We can look at it as a way of enabling one to build and use more *flexible* models in the context of *black box* modelling, where the added *flexibility* is indirectly provided by the localization properties of the wavelet transform.

In general terms, the proposed approach is based on the definition of a two-dimensional grid of time/scale over which phenomena can conceptually evolve. The observed values at the finest resolution can be seen as a result of different (eventually dynamical) structures acting on different scales, i.e., along different levels (or horocyles, Section 2.6.3) of the grid. Such dynamical relationships are established through the development of "sub-models", involving input and output variables, at a correspondent resolution (and eventually others from resolutions in the neighbourhood). These "sub-models", at each scale, can be selected from the classical arsenal, whose application scenarios are, by this way, extended to multiscale modelling situations. Doing so, will allow for the integration of the benefits associated with such techniques (accumulated experience on their use, already developed efficient algorithms, extensive theoretical support, etc.) and those related to the use of different representations of signals (multiresolution representation) that have already proven its utility in several related contexts (e.g. non-linear filtering of non-stationary signals, data compression and decorrelation of serial correlated data).

The multiscale modelling approach essentially consists on looking to collected data at different resolutions and see whether there can be any added benefit on modelling it at decomposition depths higher than zero (finest scale). If that happens to be the case, then dynamical features localized at different scales should be modelled.

More specifically, in the proposed approach, besides analysing the data collected from the system under consideration at the finest scale (corresponding to the sampling rate used for sampling), i.e., along the discrete grid of time (Figure 11.1), we also look at what is going on at higher scales, by analysing the wavelet coefficients corresponding to

higher decomposition depths, i.e., considering the two-dimensional discrete grid of time/space (Figure 11.2).

Figure 11.1. The classic discrete grid of time.

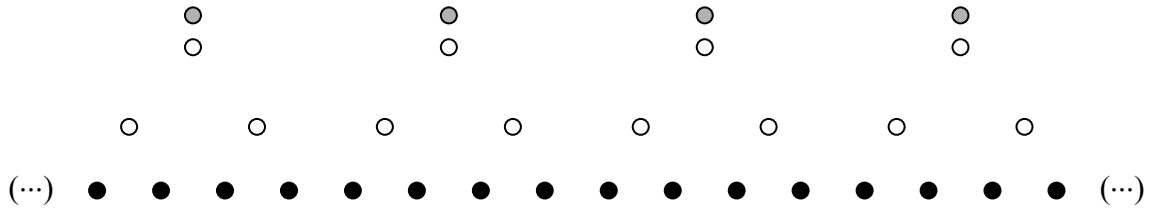


Figure 11.2. The two dimensional time/scale discrete grid, along which a process conceptually evolves in the proposed approach (white and grey points, where the white points stand for detail coefficients and the grey points for approximation coefficients; black points represents the classical grid of time). The depth of decomposition is 2. Note that the total number of points remains the same in both grids, since we are using only orthogonal, non-redundant, wavelet transforms.

As can be seen from Figure 11.2, the total number of nodes remains the same as in the classical discrete grid of time, which means that no extra information is being used, besides the one available at the finest grid, since only a transformation is being applied.

Thus, at a first stage, wavelet transform coefficients are computed, and then indexed by its nodes in the new topological structure where the process evolves (the new grid structure).

In a second stage, the "causal" *connectivity structure* of the "sub-models" is established, i.e., a decision is made regarding which are the nodes whose input and output coefficients affect the output coefficient at each "current" node. Such a recursive structure can be represented graphically in diagrams such as those in Figure 11.4, with the aid of a few conventions (Figure 11.3).

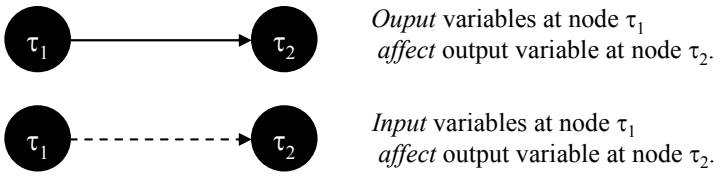


Figure 11.3. Convention for graphically representing the relationship between those nodes (where the arrow begins, τ_1) whose input and output (wavelet transformed) variables affect the (wavelet transformed) output variable at another node (where the arrow ends, τ_2).

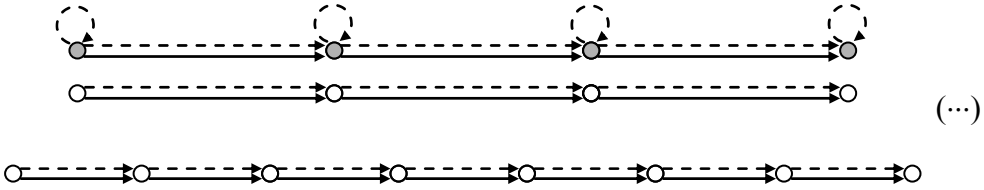


Figure 11.4. A possible multiscale dynamic recursive structure, for a decomposition depth of 2, where white points stand for detail coefficients and grey points for approximation coefficients.

In a third stage, the specific structure for all the "sub-models" is specified. To facilitate this task, we can subdivide them into separate groups, as illustrated below, using three groups of "sub-models":[42]

[42] The "sub-models" that compose the multiscale global model are grouped into three categories, to avoid an excessive use of indexing nomenclature in the equations. The first category stands for the top level nodes models, and thus refers only to the (coarsest) detail and scaling coefficients' input/ouput relationships, and naturally does not contain any dependency regarding to other coarser coefficients (because they are not calculated); the second category consists of models for the second level of output detail coefficients, and is the only one where it is allowed a dependency upon *scaling* coefficients from a coarser scale; this connective topological characteristic distinguishes the second category from the third one, where only the relationships between output detail coefficients at each node with other input or output *detail* coefficients at the same or coarser scales are considered. The second and third categories could however be fused into a single category, but at the expense of using a more cumbersome nomenclature in the description of such more general models.

1. *"Sub-models" for top level nodes* (coarser approximation and detail coefficients for $j = J_{dec}$, where J_{dec} is the decomposition depth)

$$ay(\tau) = f_{J_{dec},a}\left\{\mathbf{ay}\left[P^{ay}_{J_{dec}}(\tau)\right], \mathbf{au}\left[P^{au}_{J_{dec}}(\tau)\right]\right\} + V_a(\tau)w(\tau) \tag{11.1}$$

$$dy(\tau) = f_{J_{dec},d}\left\{\mathbf{dy}\left[P^{dy}_{J_{dec}}(\tau)\right], \mathbf{du}\left[P^{du}_{J_{dec}}(\tau)\right]\right\} + V_d(\tau)w(\tau) \tag{11.2}$$

2. *"Sub-models" for second layer nodes* (detail coefficients for $j = J_{dec} - 1$)

$$dy(\tau) = f_{J_{dec}-1}\left\{\mathbf{ay}\left[P^{ay}_{m_j(\tau)-1}(\tau)\right], \mathbf{dy}\left[P^{dy}_{m_j(\tau)-1}(\tau)\right], ..., \mathbf{du}\left[P^{du}_{J_{dec}-1}(\tau)\right]\right\} + V(\tau)w(\tau) \tag{11.3}$$

3. *"Sub-models" for lower layer nodes* (detail coefficients for $j < J_{dec} - 1$)

$$dy(\tau) = f_{m_j(\tau)}\left\{\mathbf{dy}\left[P^{dy}_{m_j(\tau)}(\tau)\right], \mathbf{du}\left[P^{du}_{m_j(\tau)}(\tau)\right]\right\} + V(\tau)w(\tau) \tag{11.4}$$

where,

$m_j(\tau)$ – provides the horocycle or scale index for node (τ);

$P(\tau)$ – represents those nodes belonging to a (properly defined) past (or causal) neighbourhood relatively to node τ (usually nodes localized in the *upper-left* quadrant, taking as reference for the origin node τ);

$P_{m_j(\tau)}$ – restriction of $P(\tau)$ to nodes belonging at the same horocycle ($m_j(\tau)$);

$ay(\tau)$, $\mathbf{ay}$ – approximation coefficient at node τ and set of approximation coefficients relative to some past horizon;

$dy(\tau)$, $\mathbf{dy}$ – detail coefficient at node τ and set of detail coefficients relative to some past horizon;

$V(\tau)$ – noise variance at node τ.

For the particular case where the "sub-models" have *linear* structures, the above general set of "sub-models" gives rise to the following general linear multiscale model structure:

1. *"Sub-models" for top level nodes*

$$ay(\tau) = \sum_{\gamma \in P^{ay}_{m_j(\tau)}(\tau)} A^{a}_{m_j(\tau),\gamma} \cdot ay(\gamma) + \sum_{\gamma \in P^{au}_{m_j(\tau)}(\tau)} B^{a}_{m_j(\tau),\gamma} \cdot au(\gamma) + V_a(\tau) w(\tau) \qquad (11.5)$$

$$dy(\tau) = \sum_{\gamma \in P^{ay}_{m_j(\tau)}(\tau)} A^{d}_{m_j(\tau),\gamma} \cdot ay(\gamma) + \sum_{\gamma \in P^{au}_{m_j(\tau)}(\tau)} B^{d}_{m_j(\tau),\gamma} \cdot au(\gamma) + V_d(\tau) w(\tau) \qquad (11.6)$$

2. *"Sub-models" for second layer nodes*

$$dy(\tau) = \sum_{\gamma \in P^{ay}_{m_j(\tau)}(\tau)} A_{m_j(\tau),\gamma} \cdot ay(\gamma) + \sum_{\gamma \in P^{dy}_{m_j(\tau)}(\tau)} A^{*}_{m_j(\tau),\gamma} \cdot dy(\gamma) + \sum_{\gamma \in P^{du}_{m_j(\tau)}(\tau)} B_{m_j(\tau),\gamma} \cdot du(\gamma) + V(\tau) w(\tau) \qquad (11.7)$$

3. *"Sub-models" for lower layer nodes*

$$dy(\tau) = \sum_{\gamma \in P^{dy}_{m_j(\tau)}(\tau)} A_{m_j(\tau),\gamma} \cdot dy(\gamma) + \sum_{\gamma \in P^{du}_{m_j(\tau)}(\tau)} B_{m_j(\tau),\gamma} \cdot du(\gamma) + V(\tau) w(\tau) \qquad (11.8)$$

At depth zero ($J_{dec} = 0$), the above multiscale model structure reduces to its classical counterpart, implemented over the discrete grid of time, but, as further decomposition depth is introduced in the analysis, the dynamics at different scales begin to be explicitly addressed. Future work should involve testing these models in systems where multiscale dynamics are known to be present and their comparison with classical approaches regarding, for instance, prediction ability.

Criteria should also be developed to assist in the selection of the relevant scales, i.e., those carrying relevant predictive information (for instance, looking at the magnitude of correlation coefficients between predictions and observations during the training phase, or through other methodologies, such as cross-validation or information theory based criteria). Only those scales that are selected during the training phase will be used in the subsequent application to fresh data.

Once the model structure is defined and estimated from available data (multiscale system identification), other tasks can be carried out, taking advantage of such a modelling formalism, namely regarding (multiscale) optimal estimation and data rectification.

11.2 Multiscale Monitoring

Alternative multiscale approaches to process monitoring can also be explored in future work, and in this subsection the foundations for another representative of this class of methods is presented, and its potential usefulness illustrated.

The proposed approach is based upon the distribution of some measure of the energy contained at the different frequency bands (indexed by the scale index) along successive non-overlapping windows of constant dyadic length (as represented in Figure 9.1-c). Coefficients from a translation invariant wavelet transformation are used for performing the calculations, in order to minimize "oscillations" in the energy contents for the different bands that can be strictly attributed to different origins established for data used in the analysis.

The approach can be used for either univariate or monitoring multivariate continuous processes operating under stationary conditions, in which case the approximation coefficients should be integrated in the analysis, or non-stationary processes, where they are discarded, with the method focused on monitoring the higher frequency bands, leaving the low frequency mode free to vary according to the non-stationary nature of the process (an example where this situation may arise is in data-driven fault detection of isolated process sensors).

The methodology for the multivariate situation is summarized in Table 11.1.

Table 11.1. Summary of the energy-based MSSPC methodology (multivariate case).

I. *Training phase*

b. Select: wavelet filter; decomposition depth (J_{dec}); whether approximation coefficients should be included in the analysis (we will assume this to be the case in what follows).

c. FOR each non-overlapping moving window of constant dyadic length ($2^{J_{dec}}$), $i = 1 : nw_{train}$, compute:

i. The translation invariant wavelet transform coefficients for all available variables ($k = 1:m$);
ii. The median of the energy[43] for each variable at each scale ($j = 1:J_{dec}+1$): $e_{i,j,k}$;

END

d. Matricize the tensor $E_{nw_{train},J_{dec}+1,m}$ (composed by element $e_{i,j,k}$ in the i,j,k entry) by keeping the dimension of time constant: $E_{nw_{train},J_{dec}+1,m} \rightarrow E_{nw_{train},(J_{dec}+1)\times m}$;

e. Choose an appropriate transformation to be applied to the columns of $E_{nw_{train},(J_{dec}+1)\times m}$, that makes its multivariate behaviour more amenable for a later implementation of SPC procedures based upon parametric probability distributions;

f. Centre and scale data: $E_{nw_{train},(J_{dec}+1)\times m} \rightarrow E^{*}_{nw_{train},(J_{dec}+1)\times m}$;

g. Compute a PCA model for $E^{*}_{nw_{train},(J_{dec}+1)\times m}$ and the statistical limits for the T^2 and Q statistics.

II. Testing phase

h. FOR each new non-overlapping moving window ($i = 1:\ldots$) compute:
   i. The translation invariant wavelet transform coefficients for all variables;
   ii. The median of the energy for each variable at each scale ($j = 1:J_{dec}+1$): $e_{i,j,k}$;
   iii. Apply transformation, as defined in the training phase;
   iv. Centre and scale data, using training phase parameters;

[43] The "energy" of a vector is here defined as the sum of squares of its components.

v. Calculate the T^2 and Q statistics and check whether there is any violation of their limits, indicating that the process is no longer operating under normal operation;

END

To illustrate this methodology, one example is presented below, regarding an univariate application.

11.2.1 An Univariate Example: Monitoring an AR(1) Process

In this example, a first order auto-regressive process is monitored using the proposed energy-based multiscale framework:

$$X(k) = \phi \cdot X(k-1) + \varepsilon(k),\ \varepsilon(k) \sim N\left(0, \sigma_\varepsilon^2\right) \tag{11.9}$$

with $\phi = 0.8$ and $\sigma_\varepsilon^2 = 4$; an additional *iid* zero-mean Gaussian noise component was also added to the data, with a standard deviation of $0.05 \times \sigma_X$.

The reference set is composed by 8192 observations, corresponding to normal operation conditions. The test set contains 16384 observations, the first half being relative to normal operation conditions, while at the beginning of the second half the autoregressive parameter was changed from 0.8 to 0.6, a value that is maintained until the final of the test set (σ_ε^2 is modified accordingly, so that the value for σ_X^2 is maintained, in order to make the change harder to be detected). Furthermore, after the observation corresponding to ¾ of the test set, another perturbation is introduced into the system, now regarding a step change of magnitude +6 (while the former perturbation in ϕ is maintained). The corresponding time series plot with 3-sigma control limits is presented in Figure 11.5, where we can see that the change that occurred at the middle of the test set (active during regions 2 and 3) passes undetected, while the effect of the set change is clearly noticed in the control chart (region 3).

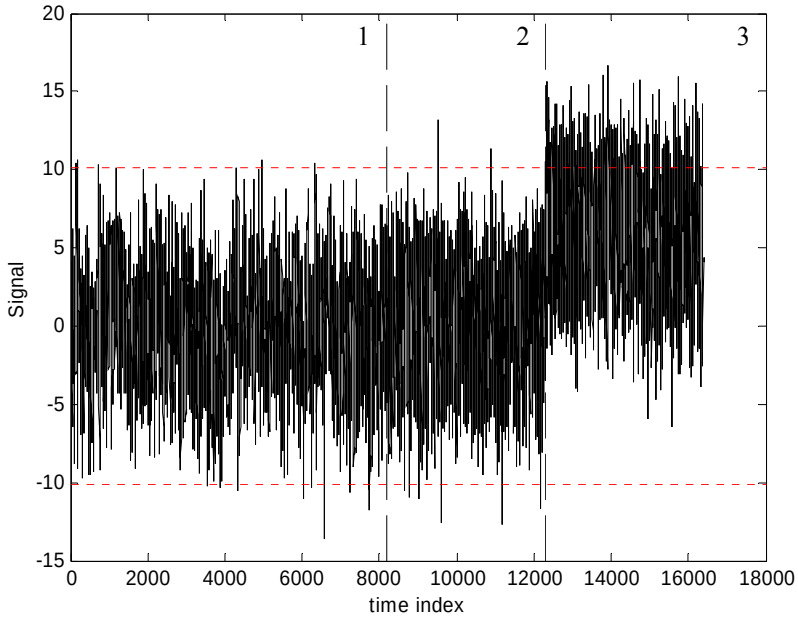


Figure 11.5. Time series plot for the test set with 3-sigma control limits. The vertical lines separate regions containing different types of testing data: 1 – normal operation; 2 – change in the autocorrelation parameter (0.8 → 0.6) and variance of the random term; 3 – step change (+6) plus the condition initiated in region 2.

Implementing the proposed energy-based multiscale approach with 3 PC's, $J_{dec} = 6$ and a power transformation ($X^{1/5}$) applied to the energy contents of detail and approximation coefficients (monitoring variables), one obtains the results presented in Figure 11.6, Figure 11.7 and Figure 11.8. In Figure 11.6 we can see that the change in the autoregressive parameter is clearly detected with both T^2 and Q statistics.

The two events that appear superimposed in region 3 can not be well resolved by the plots in Figure 11.6, but this can be appropriately done through an analysis of control charts for the principal components scores (Figure 11.7), especially looking at the behaviour of the scores for the third PC (PC3)

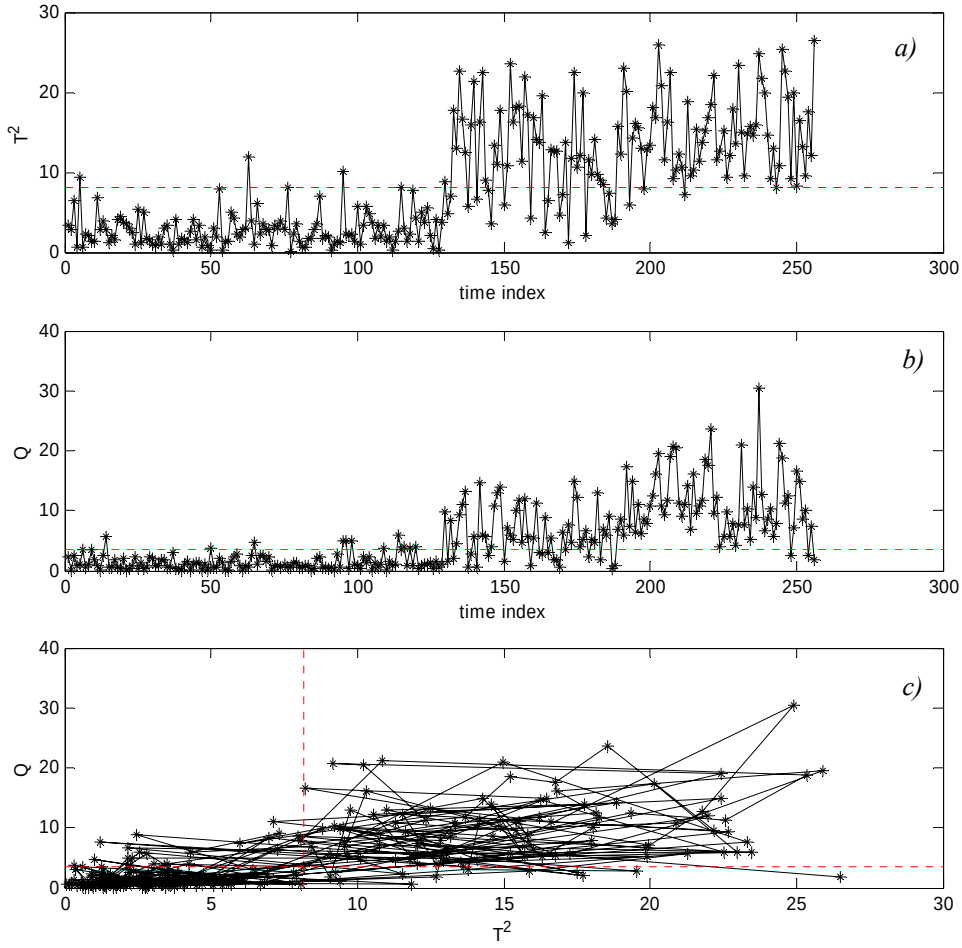


Figure 11.6. Control charts for the (a) T^2 and (b) Q statistics, plus an additional plot (c) where they are combined. Control limits for a confidence level of 95%.

The behaviour presented by the scores can be better understood by looking at the loadings for each PC (Figure 11.8), where we can see that PC1 is essentially an average of the energy distributed by the detail coefficients (first six variables), without giving much weight to the approximation coefficients (variable number 7), while the second PC is a contrast between the energy content in the details for the finest scales and coarser scales, the same happening to a lesser extent with PC3, where the approximation coefficients' energy shows now a significant importance, making this component quite sensitive to both variations in the distribution of energy across the high-medium frequency spectrum and to transitions in the operation level of the signal.

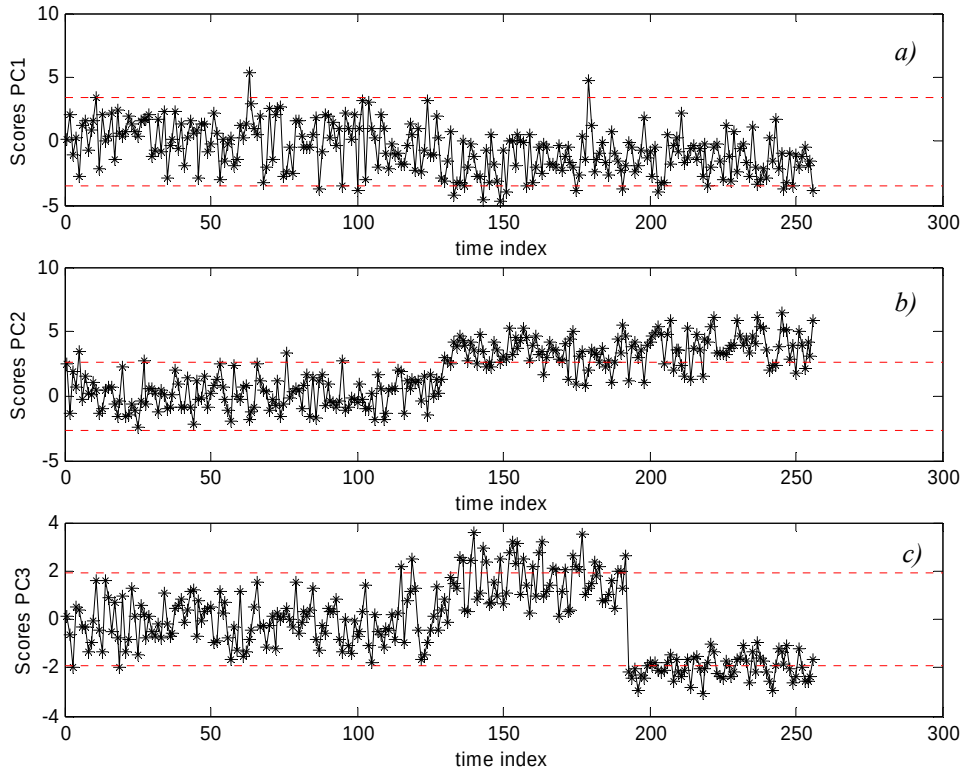


Figure 11.7. Control charts for the principal components scores: (a) PC1, (b) PC2 and (c) PC3. Control limits for a confidence level of 95%.

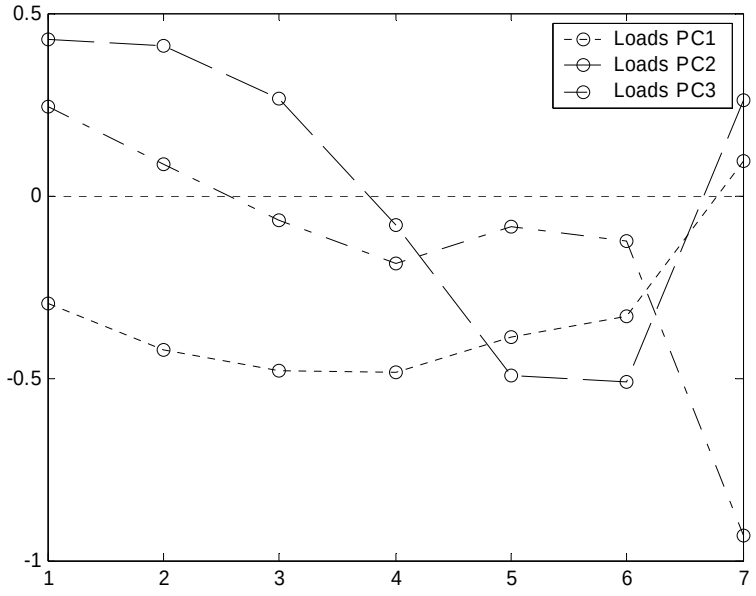


Figure 11.8. Loading vectors for the three principal components considered in the energy-based multiscale monitoring procedure.

This example illustrates the potential of the proposed methodology in monitoring even subtle dynamic upsets in the process. As future work, its properties should be better characterized and its performance assessed against other related methods.

11.3 Hierarchical Modelling of Multiresolution Networks

Chemical engineers tend to look at complex processes as pyramids where the more data and process intensive operations are carried out at the bottom of the pyramid and the higher level decisions (e.g., strategic) at the top (Saraiva, 1993). Interestingly enough, there seems to be a corresponding hierarchical structure regarding the resolution at which information is processed at the different levels of the decision-making pyramid (Figure 11.9). Operators tend to use frequent high resolution observations (minute/hour averages) to drive the process close to the desired operation level, process engineers look at summaries involving averages of hours or days to check the stability of the process, according to the current production plan, plant directors are interested in the day/month reports and administrators are more concerned with month/year figures. Each one of these elements represents a level where decisions are taken in order to comply with goals established at the higher levels.

Thus, there is a flux of information being generated at the lower levels[44] and going up the pyramid in increasingly condensed representations (lower resolutions) and decisions/goals going down the structure, affecting the operation regimes across time, as systematized in Figure 11.10, where the decision blocks (or multiresolution processing elements) receive data from the levels beneath, condense them in a suitable way and process such data in order to produce a decision that complies with the decisions/goals defined above.

[44] Information can also arise from outside the pyramid, e.g., social impact data, ecological impact figures, information about market trends and economical indicators, among others.

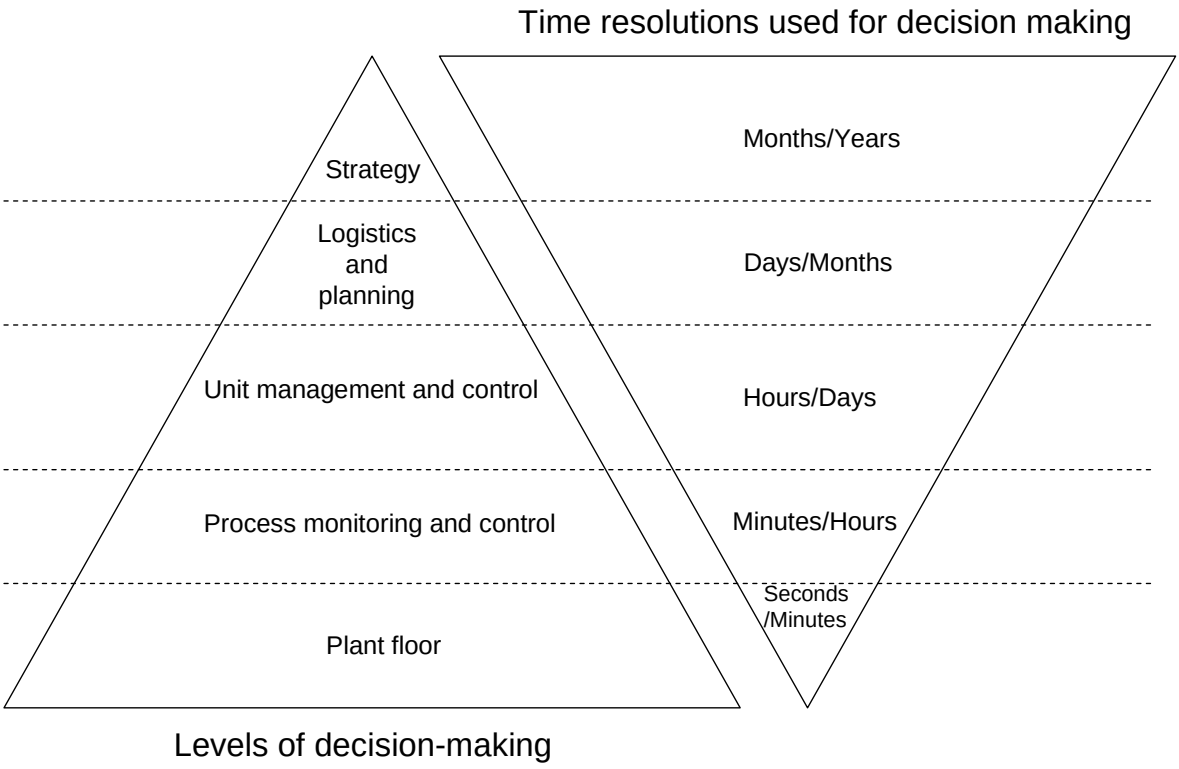


Figure 11.9. Levels of decision-making in manufacturing organizations (pyramid at the left) and the corresponding hierarchy of resolutions at which information is usually analyzed, across the different levels of decision-making (pyramid at the right).

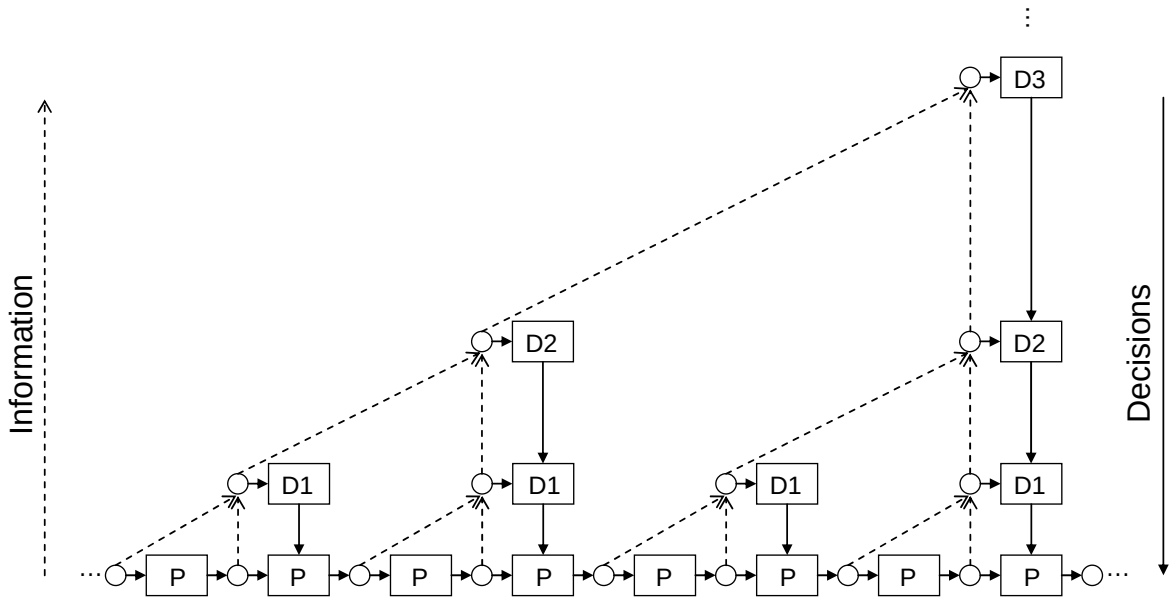


Figure 11.10. A process viewed as a hierarchical structure, where the flow of information proceeds upwards (dashed arrows) with decreasing resolution and the flow of decisions downwards (solid arrows). Each decision element analyses the condensed information derived from the lower levels, and produces a decision also targeted to these levels. Legend: P – process; Di – decision element.

11.4 Further Developments on Uncertainty-Based Methodologies

The uncertainty-based methodologies presented in Part IV-A of this thesis provide a sound way for incorporating all the available knowledge regarding data quality (values plus associated uncertainties) in the corresponding analysis. Regarding the methodologies presented here, future work should address the application of uncertainty-based regression approaches in real industrial contexts, using the guidelines extracted from the results achieved in our comparative study presented in Chapter 5. Furthermore, such a study covered a variety of data structures and noisy scenarios, but there are others interesting enough to deserve being addressed in future works, as is the case of data with correlated noise structures, specially relevant in spectroscopic applications (Wentzell & Lohnes, 1999).

The integration of uncertainty information should also be extended to *non-parametric* regression approaches. For instance, let us consider the case of (univariate) nearest neighbour regression (NNR), that consists of using only those k observations from the reference (or training) data set that are closest to the new X value (predictor), whose Y (response) we want to estimate, with the inference for Y(x) being (Hastie *et al.*, 2001):

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{11.10}$$

where $N_k(x)$ is the set of k-nearest neighbours for x, with closeness expressed in the sense of the Euclidean distance metric. When data uncertainties are also available, the distance in the X space should reflect them as well. In fact, if x is at the same Euclidean distance of x_i and x_k, but unc(x_i) > unc(x_k), it is more likely for x_i to be further way from x than x_k. Therefore, we propose the following modification of the Euclidean distance metric for the counterpart, uncertainty based approach (uNNR):

$$D_w(x_i, x_k) = \sqrt{\sum_{i=1}^{N} (x_i - x_k)^2 \left(unc(x_i)^2 + unc(x_k)^2 \right)} \tag{11.11}$$

This should be complemented with a modified version of the averaging methodology in equation (11.10), that takes also care of the uncertainty information regarding *Y*, leading to:

$$\hat{Y}(x) = \frac{\sum_{x_i \in N_{w,k}(x)} y_i / unc(y_i)^2}{\sum_{x_i \in N_{w,k}(x)} 1/unc(y_i)^2} \tag{11.12}$$

For testing this methodology, a simulation study was conducted consisting of a non-linear relationship between Y and X (a sine wave), according to the following steps:

(i) Generation of 500 samples uniformly distributed in [0,2π];

(ii) Addition of heteroscedastic noise to X and Y "true" values. The uncertainty values are randomly extracted from a uniform distribution, within a range of 0.2;

(iii) Creation of a test sample with 50 observations, and computation of the root mean square error of prediction (RMSEP) obtained for each method (NNR, uNNR).

This process was repeated 100 times for each value of the parameter "number of nearest neighbours", and the mean RMSEP values computed. As can be seen in Figure 11.11, uNNR leads in general to an improvement of prediction results. Therefore, future developments can include extending this type of uncertainty-based approach to other non-parametric regression and classification methodologies.

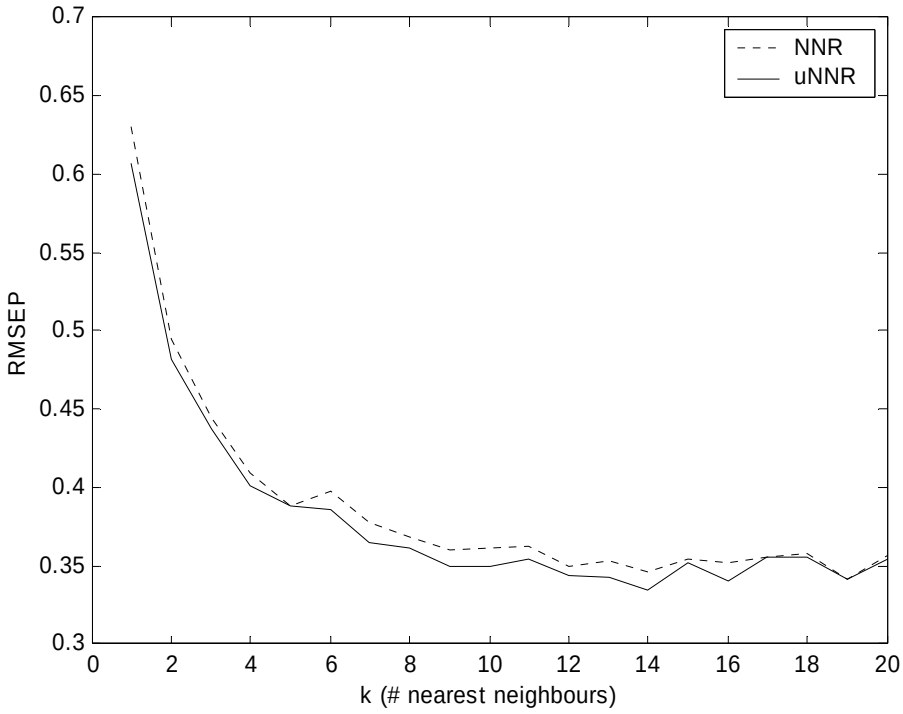


Figure 11.11. Mean RMSEP for NNR and uNNR, obtained over 100 simulations for each number of "nearest neighbours" considered (*k*).

Regarding the MRD frameworks presented in Part IV-A (Chapter 4), it is worthwhile noticing that they can be applied in rather more general contexts, other than the one explored here, where they were used to provide single-scale information (after a step of scale selection), to be processed by the tools presented in Chapters 5-7. For instance, within the scope of multiscale data analysis under the presence of missing data and uncertainty information, the uncertainty-based MRD frameworks can be integrated with MLPCA, whose formulation is well aligned with MRD frameworks.

Other scale selection approaches should also be explored in future developments, namely based on the trade-off between fitting and estimation variance of the "true" underlying signal, signal to noise ratio (SNR) measures, as well as methodologies developed for the multivariate case (e.g. exploring the scores of MLPCA and the associated uncertainty, using the tools built for the univariate case).

References

Aboufadel, E. & Schlicker, S. (1999). *Discovering Wavelets*. New York: Wiley.

Abry, P., Veitch, D. & Flandrin, P. (1998). Long-Range Dependence: Revisiting Aggregation with Wavelets. *Journal of Time Series Analysis, 19*(3), 253-266.

Al-Assaf, Y. (2004). Recognition of Control Chart Patterns Using Multi-Resolution Wavelet Analysis and Neural Networks. *Computers and Chemical Engineering, 47*, 17-29.

Alawi, A., Morris, A. J. & Martin, E. B. (2005). *Statistical Performance Monitoring Using State Space Modelling and Wavelet Analysis.* Paper presented at the ESCAPE-15, European Symposium on Computer Aided Process Engineering, Barcelona (Spain).

Alexander, S. M. & Gor, T. B. (1998). Monitoring, Diagnosis and Control of Industrial Processes. *Computers & Industrial Engineering, 35*(1-2), 193-196.

Alsberg, B. K. (1999). Multiscale Cluster Analysis. *Analytical Chemistry, 71*, 3092-3100.

Alsberg, B. K. (2000). Parsimonious Multiscale Classification Models. *Journal of Chemometrics, 14*, 529-539.

Alsberg, B. K., Woodward, A. M. & Kell, D. B. (1997). An Introduction to Wavelet Transforms for Chemometricians: A Time-Frequency Approach. *Chemometrics and Intelligent Laboratory Systems, 37*, 215-239.

Alsberg, B. K., Woodward, A. M., Winson, M. K., Rowland, J. J. & Kell, D. B. (1998). Variable Selection in Wavelet Regression Models. *Analytica Chimica Acta, 368*, 29-44.

Aradhye, H. B., Bakshi, B. R., Davis, J. F. & Ahalt, S. C. (2004). Clustering in Wavelet Domain: A Multiresolution ART Network for Anomaly Detection. *AIChE Journal, 50*(10), 2455-2466.

Aradhye, H. B., Bakshi, B. R., Strauss, R. & Davis, J. F. (2003). Multiscale SPC Using Wavelets: Theoretical Analysis and Properties. *AIChE Journal, 49*(4), 939-958.

Aradhye, H. B., Davis, J. F. & Bakshi, B. R. (2002). ART-2 and Multiscale ART-2 for On-Line Process Fault Detetction - Validation Via Industrial Case Studies and Monte Carlo Simulation. *Annual Reviews in Control, 26*, 113-127.

Arteaga, F. & Ferrer, A. (2002). Dealing with Missing Data in MSPC: Several Methods, Different Interpretations, Some Examples. *Journal of Chemometrics, 16*, 408-418.

Åström, K. J. & Eykhoff, P. (1971). System Identification – A Survey. *Automatica, 7*, 123-162.

Bakhtazad, A., Palazoglu, A. & Romagnoli, J. A. (2000). Detection and Classification of Abnormal Process Situations Using Multidimensional Wavelet Domain Hidden Markov Trees. *Computers and Chemical Engineering, 24*, 769-775.

Bakshi, B. R. (1998). Multiscale PCA with Application to Multivariate Statistical Process Control. *AIChE Journal, 44*(7), 1596-1610.

Bakshi, B. R. (1999). Multiscale Analysis and Modeling Using Wavelets. *Journal of Chemometrics, 13*, 415-434.

Bakshi, B. R., Bansal, P. & Nounou, M. N. (1997). Multiscale Rectification of Random Errors Without Process Models. *Computers and Chemical Engineering, 21*(Suppl.), S1167-S1172.

Bakshi, B. R., Koulouris, A. & Stephanopoulos, G. (1994). Learning at Multiple Resolutions: Wavelets as Basis Functions in Artificial Neural Networks and Inductive Decision Trees. In *Wavelet Applications in Chemical Engineering* (pp. 139-174). Boston: Kluwer Academic Publishers.

Bakshi, B. R. & Stephanopoulos, G. (1993). Wave-Net: a Multiresolution, Hierarchical Neural Network with Localized Learning. *AIChE Journal, 39*(1), 57-81.

Bakshi, B. R. & Stephanopoulos, G. (1996). Compression of Chemical Process Data by Functional Approximation and Feature Extraction. *AIChE Journal, 42*(2), 477-492.

Bassevile, M., Benveniste, A., Chou, K. C., Golden, S. A., Nikoukhah, R. & Willsky, A. S. (1992a). Modeling and Estimation of Multiresolution Stochastic Processes. *IEEE Transactions on Information Theory, 38*(2), 766-784.

Bassevile, M., Benveniste, A. & Willsky, A. S. (1992b). Multiscale Autoregressive Processes, Part I: Schur-Levinson Parametrizations. *IEEE Transactions on Signal Processing, 40*(8), 1915-1934.

Bassevile, M., Benveniste, A. & Willsky, A. S. (1992c). Multiscale Autoregressive Processes, Part II: Lattice Structures for Whitening and Modeling. *IEEE Transactions on Signal Processing, 40*(8), 1935-1954.

Benveniste, A., Nikoukhah, R. & Willsky, A. S. (1994). Multiscale System Theory. *IEEE Transactions on Circuits and Systems - I: Fundamental Theory and Applications, 41*(1), 2-15.

Bernardo, F. P., Pistikopoulos, E. N. & Saraiva, P. M. (1999). Integration and Computational Issues in Stochastic Design and Planning Optimization Problems. *Industrial & Engineering Chemistry Research, 38*, 3056-3068.

Bernié, J. P., Pande, H. & Gratton, R. (2004). A New Wavelet-Based Instrumental Method for Measuring Print Mottle. *Pulp & Paper Canada, 105*(9), 24-26.

Beylkin, G., Coifman, R. & Rokhlin, V. (1991). Fast Wavelet Transforms and Numerical Algorithms I. *Communications on Pure and Applied Mathematics, XLIV*, 141-183.

Bharati, M. H. & MacGregor, J. F. (1998). Multivariate Image Analysis for Real-Time Process Monitoring and Control. *Industrial & Engineering Chemistry Research, 37*, 4715-4724.

Bindal, A., Khinast, J. G. & Ierapetritou, M. G. (2003). Adaptive Multiscale Solution of Dynamical Systems in Chemical Processes Using Wavelets. *Computers and Chemical Engineering, 27*, 131-142.

Binder, T. (2002). *Adaptive Multiscale Methods for the Solution of Dynamic Optimization Problems.* PhD Thesis, RWTH Aachen University.

Bouydain, M., Colom, J. F. & Pladellorens, J. (1999). Using Wavelets to Determine Paper Formation by Light Transmission Image Analysis. *Tappi Journal, 82*(7), 153-158.

Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (1994). *Time Series Analysis - Forecasting and Control* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Braatz, R. D., Alkire, R. C., Rusli, E. & Drews, T. O. (2004). Multiscale Systems Engineering with Application to Chemical Reaction Processes. *Chemical Engineering Science, 59*, 5623-5628.

Bro, R., Sidiropoulos, N. D. & Smilde, A. K. (2002). Maximum Likelihood Fitting Using Ordinary Least Squares Algorithms. *Journal of Chemometrics, 16*, 387-400.

Bruce, A., Donoho, D. & Gao, H.-Y. (1996). Wavelet Analysis. *IEEE Spectrum*, 26-35.

Burnham, A. J., Macgregor, J. F. & Viveros, R. (1999). Latent Variable Multivariate Regression Modeling. *Chemometrics and Intelligent Laboratory Systems, 48*, 167-180.

Burrus, C. S., Gopinath, R. A. & Guo, H. (1998). *Introduction to Wavelets and Wavelet Transforms - A Primer*. New Jersey: Prentice Hall.

Carrier, J. F. & Stephanopoulos, G. (1998). Wavelet-Based Modulation in Control-Relevant Process Identification. *AIChE Journal, 44*(2), 341-360.

Carvalho, M. G. V., Martins, A. A. & Figueiredo, M. M. L. (2003). Kraft Pulpling of Portuguese Eucalyptus globulus: Effect of Process Conditions on Yield and Pulp Properties. *Appita*, 267-274.

Chan, Y. T. (1995). *Wavelet Basics*. Boston: Kluwer.

Charpentier, J. C. & McKenna, T. F. (2004). Managing Complex Systems: Some Trends for the Future of Chemical and Process Engineering. *Chemical Engineering Science, 59*, 1617-1640.

Chatfield, C. (1989). *The Analysis of Time Series - An Introduction* (4th ed.). London: Chapman and Hall.

Chau, F.-T., Liang, Y.-Z., Gao, J. & Shao, X.-G. (2004). *Chemometrics - From Basics to Wavelet Transform* (Vol. 164). Hoboken, NJ: Wiley.

Chen, B. H., Wang, X. Z., Yang, S. H. & McGreavy, C. (1999). Application of Wavelets and Neural Networks to Diagnostic System Development, 2, an Integrated Framework and its Application. *Computers and Chemical Engineering, 23*, 945-954.

Chiang, L. H., Russel, E. L. & Braatz, R. D. (2001). *Fault Detection and Diagnosis in Industrial Systems*. London: Springer-Verlag.

Chou, K. C. (1991). *A Stochastic Modeling Approach to Multiscale Signal Processing.* PhD Thesis, MIT.

Chou, K. C., Golden, S. A. & Willsky, A. S. (1993). Multiresolution Stochastic Models, Data Fusion, and Wavelet Transforms. *Signal Processing, 34*, 257-282.

Chou, K. C., Willsky, A. S. & Benveniste, A. (1994a). Multiscale Recursive Estimation, Data Fusion, and Regularization. *IEEE Transactions on Automatic Control, 39*(3), 464-478.

Chou, K. C., Willsky, A. S. & Nikoukhah, R. (1994b). Multiscale Systems, Kalman Filters, and Riccati Equations. *IEEE Transactions on Automatic Control, 39*(3), 479-492.

Chui, C. K. (1992). *An Introduction to Wavelets*. San Diego (CA, USA): Academic Press.

Chui, C. K. & Chen, G. (1999). *Kalman Filtering with Real-Time Applications* (3rd ed. Vol. 17). Berlin: Springer.

Claus, B. (1993). Multiscale Statistical Signal Processing: Identification of a Multiscale AR Process from a Sample of an Ordinary Signal. *IEEE Transactions on Signal Processing, 41*(12), 3266-3274.

Cocchi, M., Durante, C., Foca, G., Manzini, D., Marchetti, A. & Ulrici, A. (2004). Application of a Wavelet-Based Algorithm on HS-SPME/GC Signals for the Classification of Balsamic Vinegars. *Chemometrics and Intelligent Laboratory Systems, 71*, 129-140.

Cocchi, M., Seeber, R. & Ulrici, A. (2001). WPTER: Wavelet Packet Transform for Efficient Pattern Recognition of Signals. *Chemometrics and Intelligent Laboratory Systems, 57*, 97-119.

Cocchi, M., Seeber, R. & Ulrici, A. (2003). Multivariate Calibration of Analytical Signals by WILMA (Wavelet Interface to Linear Modelling Analysis). *Journal of Chemometrics, 17*, 512-527.

Coelho, C. J., Galvão, R. K. H., Araujo, M. C. U., Pimentel, M. F. & Da Silva, E. C. (2003). A Solution to the Wavelet Transform Optimization Problem in Multicomponent Analysis. *Chemometrics and Intelligent Laboratory Systems, 66*, 205-217.

Cohen, A. & Ryan, R. D. (1995). *Wavelets and Multiscale Signal Processing*. London: Chapman & Hall.

Cohen, I., Raz, S. & Malah, D. (1999). Translation-Invariant Denoising Using the Minimum Description Length Criterion. *Signal Processing, 75*, 201-223.

Coifman, R. R. & Donoho, D. L. (1995). *Translation-Invariant De-Noising* (Technical report): Department of Statistics, Stanford University.

Coifman, R. R. & Wickerhauser, M. V. (1992). Entropy-Based Algorithms for Best Basis Selection. *IEEE Transactions on Information Theory, 38*(2), 713-718.

Costa, R., Angélico, D., Reis, M. S., Ataíde, J. & Saraiva, P. M. (2005). Paper Superficial Waviness: Conception and Implementation of an Industrial Statistical Measurement System. *Analytica Chimica Acta, 544*, 135-142.

Costa, R., Angélico, D., Saraiva, P., Reis, M., Ataíde, J., Abreu, C., et al. (2004). *Paper Superficial Waviness: Conception of a Statistical Measurement System.* Paper presented at the CAC-2004 Chemometrics in Analytical Chemistry, Lisbon (Portugal).

Crato, N. (1998, 19/09/1998). A "Ondinha" do FBI. *Expresso,* pp. 132-133.

Crouse, M. S., Nowak, R. D. & Baraniuk, R. G. (1998). Wavelet-Based Statistical Signal Processing Using Hidden Markov Models. *IEEE Transactions on Signal Processing, 46*(4), 886-902.

Dable, B. K. & Booksh, K. S. (2001). Selecting Significant Factors by the Noise Addition Method in Principal Components Analysis. *Journal of Chemometrics, 15*, 591-613.

Dai, X. D., Motard, R. L., Joseph, B. & Silverman, D. C. (2000). Corrosion Process Monitoring Using Wavelet Analysis. *Industrial & Engineering Chemistry Research, 39*, 1256-1263.

Daiguji, M., Kudo, O. & Wada, T. (1997). Application of Wavelet Analysis to Fault Detection in Oil Refinery. *Computers and Chemical Engineering, 21*(Suppl.), S1117-S1122.

Daoudi, K., Frakt, A. B. & Willsky, A. S. (1999). Multiscale Autoregressive Models and Wavelets. *IEEE Transactions on Information Theory, 45*(3), 828-845.

Daubechies, I. (1992). *Ten Lectures on Wavelets* (Vol. 61). Philadelphia, Pennsylvania: SIAM.

De Castro, M., Galea-Rojas, M., Bolfarine, H. & De Castilho, M. V. (2004). Detection of Analytical Bias When Comparing Two or More Measuring Methods. *Journal of Chemometrics, 18*, 431-440.

de Jong, S., Wise, B. W. & Ricker, N. L. (2001). Canonical Partial Least Squares and Continuum Power Regression. *Journal of Chemometrics, 15*, 85-100.

Depczynsky, U., Jetter, K., Molt, K. & Niemöller, A. (1999). The Fast Wavelet Transform on Compact Intervals as a Tool in Chemometrics - II. Boundary Effects, Denoising and Compression. *Chemometrics and Intelligent Laboratory Systems, 49*, 151-161.

Dijkerman, R. W. & Mazumdar, R. R. (1994). Wavelet Representations of Stochastic Processes and Multiresolution Stochastic Models. *IEEE Transactions on Signal Processing, 42*(7), 1640-1652.

Donald, D., Everingham, Y. & Coomans, D. (2005). Integrated Wavelet Principal Component Mapping for Unsuppervised Clustering on Near Infra-Red Spectra. *Chemometrics and Intelligent Laboratory Systems, 77*, 32-42.

Donoho, D. L. (1995). De-Noising by Soft-Thresholding. *IEEE Transactions on Information Theory, 41*(3), 613-627.

Donoho, D. L. & Johnstone, I. M. (1992). *Ideal Spatial Adaptation by Wavelet Shrinkage* (Technical report): Department of Statistics, Stanford University.

Doroslovački, M. I. & Fan, H. (1996). Wavelet-Based Linear System Modeling and Adaptative Filtering. *IEEE Transactions on Signal Processing, 44*(5), 1156-1167.

Doroslovački, M. I., Fan, H. & Yao, L. (1998). Wavelet-Based Identification of Linear Discrete-Time Systems: Robustness Issue. *Automatica, 34*(12), 1627-1640.

Doymaz, F., Bakhtazad, A., Romagnoli, J. A. & Palazoglu, A. (2001). Wavelet-Based Robust Filtering of Process Data. *Computers and Chemical Engineering, 25*, 1549-1559.

Draper, N. R. & Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). NY: Wiley.

Duin, R. P. W., Juszczak, P., Paclik, P., Pekalska, E., de Ridder, D. & Tax, D. M. J. (2004). *PRTools4, A Matlab Toolbox for Pattern Recognition*: Delft University of Technology.

Dyer, M. (2000). *A Multiscale Approach to State Estimation with Applications in Process Operability Analysis and Model Predictive Control.* PhD Thesis, MIT.

Eriksson, L., Johansson, E., Kettaneh-Wold, N. & Wold, S. (2001). *Multi- and Megavariate Data Analysis – Principles and Applications*. Umeå (Sweden): Umetrics AB.

Eriksson, L., Trygg, J., Johansson, E., Bro, R. & Wold, S. (2000). Orthogonal Signal Correction, Wavelet Analysis, and Multivariate Calibration of Complicated Process Fluorescence Data. *Analytica Chimica Acta, 420*, 181-195.

Estienne, F., Pasti, L., Centner, V., Walczak, B., Despagne, F., Rimbaud, D. J., et al. (2001). A Comparison of Multivariate Calibration Techniques Applied to Experimental NIR Data Sets. Part II Predictive Ability Under Extrapolation Conditions. *Chemometrics and Intelligent Laboratory Systems, 58*, 195-211.

Faber, K. (2000). Comparison of Two Recently Proposed Expressions for Partial Least Squares Regression Prediction Error. *Chemometrics and Intelligent Laboratory Systems, 52*, 123-134.

Faber, K. & Bro, R. (2002). Standard Error of Prediction for Multiway PLS: 1. Background and a Simulation Study. *Chemometrics and Intelligent Laboratory Systems, 61*(133-149).

Faber, K. & Kowalski, B. R. (1997). Propagation of Measurement Errors for the Validation of Predictions Obtained by Principal Component Regression and Partial Least Squares. *Journal of Chemometrics, 11*, 181-238.

Fieguth, P. W. & Willsky, A. S. (1996). Fractal Estimation Using Models on Multiscale Trees. *IEEE Transactions on Signal Processing, 44*(5), 1297-1300.

Flandrin, P. (1989). On the Spectrum of Fractional Brownian Motion. *IEEE Transactions on Information Theory, 35*(1), 197-199.

Flandrin, P. (1992). Wavelet Analysis and Synthesis of Fractional Brownian Motion. *IEEE Transactions on Information Theory, 38*(2), 910-917.

Flores-Cerrilo, J. & MacGregor, J. F. (2004). Control of Batch Product Quality by Trajectory Manipulation Using Latent Variable Models. *Journal of Process Control, 14*, 539-553.

Forseth, T. & Helle, T. (1996). *Principles and Application of a Method for Assessing Detailed Paper Surface Topography.* Paper presented at the 82nd Annual Meeting, Technical Session, CPPA.

Fourie, S. H. & de Vaal, P. (2000). Advanced Process Monitoring Using an On-Line Non-Linear Multiscale Principal Component Analysis Methodology. *Computers and Chemical Engineering, 24*, 755-760.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. San Diego: Academic Press.

Gabrielsson, J., Lindberg, N.-O. & Lundstedt, T. (2002). Multivariate Methods in Pharmaceutical Applications. *Journal of Chemometrics, 16*, 141-160.

Galea-Rojas, M., de Castilho, M. V., Bolfarine, H. & de Castro, M. (2003). Detection of Analytical Bias. *Analyst, 128*, 1073-1081.

Galvão, R. K. H., José, G. E., Filho, H. A. D., Araujo, M. C. U., da Silva, E. C., Paiva, H. M., et al. (2004). Optimal Wavelet Filter Construction Using X and Y Data. *Chemometrics and Intelligent Laboratory Systems, 70*, 1-10.

Ganesan, R., Das, T. K. & Venkataraman, V. (2004). Wavelet Based Multiscale Process Monitoring - A Literature Review. *IIE Transactions on Quality and Reliability Engineering, 36*(9), 787-806.

Geladi, P. & Kowalski, B. R. (1986). Partial Least-Squares Regression: a Tutorial. *Analytica Chimica Acta, 185*, 1-17.

Golden, S. A. (1991). *Identifying Multiscale Statistical Models Using the Wavelet Transform.* MSc Thesis, MIT.

Golub, G. H. & Van Loan, C. F. (1989). *Matrix Computations*. Baltimore, London: The John Hopkins University Press.

Goodman, L. A. & Haberman, S. J. (1990). The Analysis of Nonadditivity in Two-way Analysis of Variance. *Journal of the American Statistical Association, 85*(409), 109-135.

Guimarães, R. C. & Cabral, J. A. S. (1997). *Estatística*. Lisboa: McGraw-Hill.

Haaland, D. M. & Thomas, E. V. (1988). Partial Least-Squares Methods for Spectral Analysis. 1. Relation to Other Quantitative Calibration Methods and the Extraction of Qualitative Information. *Analytical Chemistry, 60*, 1193-1202.

Hasiewicz, Z. (1999). Hammerstein System Identification by the Haar Multiresolution Approximation. *International Journal of Adaptive Control and Signal Processing, 13*, 691-717.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*. NY: Springer.

Helland, I. S. (1988). On the Structure of Partial Least Squares Regression. *Commun. Statist.-Simula., 17*(2), 581.

Helland, I. S. (2001a). *Rotational Symmetry, Model Reduction and Optimality of Prediction from the PLS Population Model.* Paper presented at the 2nd International Symposium on PLS and Related Methods.

Helland, I. S. (2001b). Some Theoretical Aspects of Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems, 58*, 97-107.

Helland, I. S. (2002). Partial Least Squares Regression. In *Encyclopedia of Statistical Sciences* (2nd ed.).

Herrador, M. Á., Asuero, A. G. & González, A. G. (2005). Estimation of the Uncertainty of Indirect Measurements from the Propagation of Distributions by Using the Monte-Carlo Method: An Overview. *Journal of Chemometrics, 79*, 115-122.

Herrick, D. R. M. (2000). *Wavelet Methods for Curve and Surface Estimation.* PhD Thesis, University of Bristol.

Ho, T. T. (1998). *Multiscale Modeling and Estimation of Large-Scale Dynamic Systems.* PhD Thesis, MIT.

Höskuldsson, A. (1988). PLS Regression Methods. *Journal of Chemometrics, 2*, 211-228.

Höskuldsson, A. (1996). *Prediction Methods in Science and Technology*. Thor Publishing.

Hubbard, B. B. (1998). *The World According to Wavelets - The Story of a Mathematical Technique in the Making* (2nd ed.). Natick, Massachusetts: A K Peters.

Hunter, J. S. (1986). The Exponentially Weighted Moving Average. *Journal of Quality Technology, 18*(4), 203-210.

Indahl, U. G. & Naes, T. (1998). Evaluation of Alternative Spectral Feature Extraction Methods of Textural Images for Multivariate Modeling. *Journal of Chemometrics, 12*, 261-278.

ISO. (1993). *Guide to the Expression of Uncertainty*. Geneva, Switzerland.

ISO. (1996). ISO 11562: 1996. Geometrical Product Specifications (GPS) - Surface texture: Metrological characteristics of phase correct filters.

ISO. (1997). ISO 4287: 1997. Geometrical Product Specifications (GPS) - Surface texture: Profile method - terms, definitions and surface texture parameters.

Izadi, I., Zhao, Q. & Chen, T. (2005). An Optimal Scheme for Fast Rate Fault Detection Based on Multirate Sampled Data. *Journal of Process Control, 15*, 307-319.

Jackson, J. E. (1959). Quality Control Methods for Several Related Variables. *Technometrics, 1*(4), 359-377.

Jackson, J. E. (1991). *A User's Guide to Principal Components*. New York: Wiley.

Jackson, J. E. & Mudholkar, G. S. (1979). Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics, 21*(3), 341-349.

Jaeckle, C. & MacGregor, J. F. (1998). Product Design through Multivariate Statistical Analysis of Process Data. *AIChE Journal, 44*(5), 1105-1118.

Jaeckle, C. & MacGregor, J. F. (2000). Product Transfer Between Plants Using Historical Process Data. *AIChE Journal, 46*(10), 1989-1997.

Jansen, M. (2001). *Noise Reduction by Wavelet Thresholding* (Vol. 161). NY: Springer.

Jazwinsky, A. H. (1970). *Stochastic Processes and Filtering Theory*. New York: Academic Press.

Jeong, M. K., Lu, J.-C., Huo, X., Vidakovic, B. & Chen, D. (2004). *Wavelet-Based Data Reduction Techniques for Process Fault Detection* (No. 11/2004): Georgia Institute of Technology, School of Industrial and Systems Engineering, Statistics Group.

Jiang, T., Chen, B. & He, X. (2000). Industrial Application of Wavelet Transform to the On-Line Prediction of Side Draw Qualities of Crude Unity. *Computers and Chemical Engineering, 24*, 507-512.

Jiao, X. J., Davies, M. S. & Dumont, G. A. (2004). Wavelet Packet Analysis of Paper Machine Data for Control Assessment and Trim Loss Optimization. *Pulp & Paper Canada, 105*(9), 34-37.

Jin, J. & Shi, J. (1999). Feature-Preserving Data Compression of Stamping Tonnage Information Using Wavelets. *Technometrics, 41*(4), 327-339.

Jin, L. & Shi, J. (2001). Automatic Feature Extraction of Waveform Signals for In-Process Diagnostic Peformance Improvement. *Journal of Intelligent Manufacturing, 12*, 257-268.

Jobson, J. D. (1992). *Applied Multivariate Data Analysis* (Vol. 2: Categorical and Multivariate Methods). New York: Springer-Verlag.

Johnson, R. A. & Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis* (3rd ed.). Prentice Hall.

Johnson, R. A. & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis* (5th ed.). Prentice Hall.

Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). New York: Springer.

Jouan-Rimbaud, D., Walczak, B., Poppi, R. J., de Noord, O. E. & Massart, D. L. (1997). Application of Wavelet Transform to Extract the Relevant Component from Spectral Data for Multivariate Calibration. *Analytical Chemistry, 69*, 4317-4323.

Juditsky, A., Zhang, Q., Delyon, B., Glorennec, P.-Y. & Benveniste, A. (1994). *Wavelets in Identification - wavelets, splines, neurons, fuzzies: how good for identification?* (Technical Report No. 2315): INRIA.

Kaiser, G. (1994). *A Friendly Guide to Wavelets*. Boston: Birkhäuser.

Kajanto, I., Laamanen, J. & Kainulainen, M. (1998). Paper Bulk and Surface. In K. Niskanen (Ed.), *Paper Physics* (Vol. 16, pp. 88-113). Jyväskylä, Finland: Fapet Oy.

Kang, L. & Albin, S. L. (2000). On-Line Monitoring When the Process Yields a Linear Profile. *Journal of Quality Technology, 32*(4), 418-426.

Kano, M., Nagao, K., Hasebe, S., Hashimoto, I., Ohno, H., Strauss, R., et al. (2002). Comparison of Multivariate Statistical Process Monitoring Methods with Applications to the Eastman Challenge Problem. *Computers and Chemical Engineering, 26*, 161-174.

Kapoor, S. G. & Wu, S. M. (1978). A Stochastic Approach to Paper Surface Characterization and Printability Criteria. *Journal of Physics D: Applied Physics, 11*, 83-96.

Kapoor, S. G. & Wu, S. M. (1979). Predicting Printability of Gravure and Embossed Papers by Time Series Analysis. *Journal of Physics D: Applied Physics, 12*, 2005-2017.

Karsligil, O. (2000). *Multiscale Modeling and Model Predictive Control of Processing Systems.* PhD Thesis, MIT.

Kaspar, M. H. & Ray, W. H. (1993a). Dynamic PLS Modelling for Process Control. *Chemical Engineering Science, 48*(20), 3447-3461.

Kaspar, M. H. & Ray, W. H. (1993b). Partial Least Squares Modelling as Successive Singular Value Decompositions. *Computers and Chemical Engineering, 17*(10), 985-989.

Kendall, M., Stuart, A. & Ord, J. K. (1983). *The Advanced Theory of Statistics* (4th ed. Vol. 3). London: Charles Griffin & Company Limited.

Kenett, R. S. & Zacks, S. (1998). *Modern Industrial Statistics - Design and Control of Quality and Reliability*. Pacific Grove: Duxbury Press.

Kessel, W. (2002). Measurement Uncertainty According to ISO/BIPM-GUM. *Thermochimica Acta, 382*, 1-16.

Kim, K., Mahmoud, M. A. & Woodall, W. H. (2003). On the Monitoring of Linear Profiles. *Journal of Quality Technology, 35*(3), 317-328.

Kimothi, S. K. (2002). *The Uncertainty of Measurements*. Milwaukee: ASQ.

Kortschot, M. T. (1997). *The Role of the Fibre in the Structural Hierarchy of Paper.* Paper presented at the The Fundamentals of Papermaking Materials - Transactions of the 11th Fundamental Research Symposium.

Kosanovich, K. A., Moser, A. R. & Piovoso, M. J. (1995). Poisson Wavelets Applied to Model Identification. *Journal of Process Control, 5*(4), 225-234.

Kosanovich, K. A. & Piovoso, M. J. (1997). PCA of Wavelet Transformed Process Data for Monitoring. *Intelligent Data Analysis, 1*(1-4), 85-99.

Kourti, T. & MacGregor, J. F. (1995). Process Analysis, Monitoring and Diagnosis, Using Multivariate Projection Methods. *Chemometrics and Intelligent Laboratory Systems, 28*, 3-21.

Kourti, T. & MacGregor, J. F. (1996). Multivariate SPC Methods for Process and Product Monitoring. *Journal of Quality Technology, 28*(4), 409-428.

Kresta, J. V., MacGregor, J. F. & Marlin, T. E. (1991). Multivariate Statistical Monitoring of Process Operating Performance. *The Canadian Journal of Chemical Engineering, 69*, 35-47.

Kresta, J. V., Marlin, T. E. & MacGregor, J. F. (1994). Development of Inferential Process Models Using PLS. *Computers and Chemical Engineering, 18*, 597-611.

Kreyszig, E. (1978). *Introductory Functional Analysis with Applications*. New York: Wiley.

Krishnan, A. & Hoo, K. A. (1999). A Multiscale Model Predictive Control Strategy. *Industrial & Engineering Chemistry Research, 38*, 1973-1986.

Krzanowski, W. J. (1979). Between-Groups Comparison of Principal Components. *Journal of the American Statistical Association, 74*(367), 703-707.

Ku, W., Storer, R. H. & Georgakis, C. (1995). Disturbance Detection and Isolation by Dynamic Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems, 30*, 179-196.

Lada, E. K., Lu, J.-C. & Wilson, J. R. (2002). A Wavelet-Based Procedure for Process Fault Detection. *IEEE Transactions on Semiconductor Manufacturing, 15*(1), 79-90.

Lakshminarayanan, S., Shah, S. L. & Nandakumar, K. (1997). Modeling and Control of Multivariable Processes: Dynamic PLS Approach. *AIChE Journal, 43*(9), 2307-2322.

Leung, A. K.-M., Chau, F.-T. & Gao, J.-B. (1998a). A Review of Applications of Wavelet Transform Techniques in Chemical Analysis. *Chemometrics and Intelligent Laboratory Systems, 43*, 165-184.

Leung, A. K.-M., Chau, F.-T., Gao, J.-B. & Shih, T.-M. (1998b). Application of Wavelet Transform in Infrared Spectrometry: Spectral Compression and Library Search. *Chemometrics and Intelligent Laboratory Systems, 43*, 69-88.

Li, J. & Kwauk, M. (2003). Exploring Complex Systems in Chemical Engineering - The Multi-Scale Methodology. *Chemical Engineering Science, 58*, 521-535.

Li, M. & Christofides, P. D. (2005). Multi-Scale Modeling and Analysis of an Industrial HVOF Thermal Spray Process. *Chemical Engineering Science, 60*, 3649-3669.

Li, X. & Qian, Y. (2004). *Process Monitoring Based on Nonlinear Wavelet Packet Principal Component Analysis.* Paper presented at the ESCAPE-14, European Symposyum on Computer Aided Process Engineering, Lisbon (Portugal).

Li, X., Qian, Y., Huang, Q. & Jiang, Y. (2004). *Multi-Scale ART2 for State Identification of Process Operation Systems.* Paper presented at the Process Systems Engineering 2003, Kunming (China).

Lira, I. (2002). *Evaluating the Measurement Uncertainty*. Bristol: Institute of Physics Publishing.

Lisý, J. M., Cholvadová, A. & Kutej, J. (1990). Multiple Straight-line Least Squares Analysis with Uncertainties in all Variables. *Computers & Chemistry, 14*, 189-192.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd ed.). Hoboken (NJ): Wiley.

Liu, G. P., Billings, S. A. & Kadirkamanathan, V. (2000a). Nonlinear System Identification using Wavelet Networks. *International Journal of Systems Science, 31*(12), 1531-1541.

Liu, Y., Cameron, I. T. & Wang, F. Y. (2000b). The Wavelet-Collocation Method for Transient Problems with Steep Gradients. *Chemical Engineering Science, 55*, 1729-1734.

Ljung, L. (1999). *System Identification - Theory for the User* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Lopes, J. P. M. A. (2001). *Supervisão e Diagnóstico de Processos Farmacêuticos com Métodos Inteligentes de Análise de Dados.* PhD Thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisbon, Portugal.

Louis, A. K., Maab, P. & Rieder, A. (1997). *Wavelets – Theory and Applications*. Chichester: Wiley.

Lu, N., Yang, Y., Gao, F. & Wang, F. (2004). Multirate Dynamic Inferential Modeling for Multivariable Processes. *Chemical Engineering Science, 59*, 855-864.

Luettgen, M. R., Karl, W. C., Willsky, A. S. & Tenney, R. R. (1993). Multiscale Representations of Markov Random Fields. *IEEE Transactions on Signal Processing, 41*(12), 3377-3396.

Luo, R., Misra, M. & Himmelblau, D. M. (1999). Sensor Fault Detection via Multiscale Analysis and Dynamic PCA. *Industrial & Engineering Chemistry Research, 38*, 1489-1495.

Luo, R., Misra, M., Qin, S. J., Barton, R. & Himmelblau, D. M. (1998). Sensor Fault Detection via Multiscale Analysis and Nonparametric Statistical Inference. *Industrial & Engineering Chemistry Research, 37*, 1024-1032.

Luyben, W. L. (1990). *Process Modeling, Simulation and Control for Chemical Engineering* (2nd ed.). New York: McGraw-Hill.

MacGregor, J. F., Jaeckle, C., Kiparissides, C. & Koutoudi, M. (1994). Process Monitoring and Diagnosis by Multiblock PLS Methods. *AIChE Journal, 40*(5).

MacGregor, J. F. & Kourti, T. (1995). Statistical Process Control of Multivariate Processes. *Control Engineering Practice, 3*(3), 403-414.

MacGregor, J. F. & Kourti, T. (1998). *Multivariate Statistical Treatment of Historical Data for Productivity and Quality Improvements.* Paper presented at the Foundation of Computer Aided Process Operations - FOCAPO 98.

MacGregor, M. A. (2001). Some Impacts of Paper Making on Paper Structure. *Paper Technology, 42*(3), 30-44.

Magnus, J. R. & Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Chichester: Wiley.

Mahadevan, N. & Hoo, K. A. (2000). Wavelet-Based Model Reduction of Distributed Parameter Systems. *Chemical Engineering Science, 55*, 4271-4290.

Mallat, S. (1989). A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 11*(7), 674-693.

Mallat, S. (1998). *A Wavelet Tour of Signal Processing*. San Diego: Academic Press.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing* (2nd ed.). San Diego: Academic Press.

Mandel, J. (1964). *The Statistical Analysis of Experimental Data*. New York: Wiley.

Martens, H. & Mevik, B.-H. (2001). Understanding the Collinearity Problem in Regression and Discriminant Analysis. *Journal of Chemometrics, 15*, 413-426.

Martens, H. & Naes, T. (1989). *Multivariate Calibration*. Chichester: Wiley.

Martin, E. B. & Morris, A. J. (1996). Non-Parametric Confidence Bounds for Process Performance Monitoring Charts. *Journal of Process Control, 6*(6), 349-358.

Martínez, À., Riu, J. & Rius, F. X. (2000). Lack of Fit in Linear Regression Considering Errors in both Axis. *Chemometrics and Intelligent Laboratory Systems, 54*, 61-73.

Martínez, À., Riu, J. & Rius, F. X. (2002a). Application of the Multivariate Least Squares Regression Method to PCR and Maximum Likelihood PCR Techniques. *Journal of Chemometrics, 16*, 189-197.

Martínez, À., Riu, J. & Rius, F. X. (2002b). Evaluating Bias in Method Comparison Studies Using Linear Regression with Errors in Both Axes. *Journal of Chemometrics, 16*, 41-53.

Masry, E. (1993). The Wavelet Transform of Stochastic Processes with Stationary Increments and Its Application to Fractional Brownian Motion. *IEEE Transactions on Information Theory, 39*(1), 260-264.

Maulud, A. H. S., Wang, D. & Romagnoli, J. A. (2005, 29 May - 1 June). *Wavelet-Based Nonlinear Multivariate Statistical Process Control.* Paper presented at the ESCAPE-15, European Symposium on Computer Aided Process Engineering, Barcelona (Spain).

Meloun, M., Čapek, J., Mikšík, P. & Brereton, R. G. (2000). Critical Comparison of Methods Predicting the Number of Components in Spectroscopic Data. *Analytica Chimica Acta, 423*, 51-68.

Meyer, Y. & Ryan, R. D. (1993). *Wavelets - Algorithms & Applications*. Philadelphia: SIAM.

Misra, M., Kumar, S., Qin, S. J. & Seemann, D. (2001). Error Based Criterion for On-Line Wavelet Data Compression. *Journal of Process Control, 11*, 717-731.

Misra, M., Qin, S. J., Kumar, S. & Seemann, D. (2000). On-Line Data Compression and Error Analysis Using Wavelet Technology. *AIChE Journal, 46*(1), 119-132.

Misra, M., Yue, H. H., Qin, S. J. & Ling, C. (2002). Multivariate Process Monitoring and Fault Diagnosis by Multi-Scale PCA. *Computers and Chemical Engineering, 26*, 1281-1293.

Mittermayr, C. R., Lendl, B., Rosenberg, E. & Grasserbauer, M. (1999). The Application of the Wavelet Power Spectrum to Detect and Estimate 1/f Noise in the Presence of Analytical Signals. *Analytica Chimica Acta, 388*, 303-313.

Montgomery, D. C. (2001). *Introduction to Statistical Quality Control* (4th ed.). New York: Wiley.

Montgomery, D. C. & Runger, G. C. (1999). *Applied Statistics and Probability for Engineers* (2nd ed.). New York: Wiley.

Motard, R. L. & Joseph, B. (Eds.). (1994). *Wavelet Applications in Chemical Engineering*: Kluwer.

Naes, T., Isaksson, T., Fearn, T. & Davies, T. (2002). *A User-Friendly Guide to Multivariate Calibration and Classification*. Chichester (UK): NIR Publications.

Naes, T. & Mevik, B.-H. (2001). Understanding the Collinearity Problem in Regression and Discriminant Analysis. *Journal of Chemometrics, 15*, 413-436.

Nason, G. P. (1994). *Wavelet Regression by Cross-Validation* (Technical Report No. 447): Department of Statistics, Stanford University.

Nason, G. P. (1995a). Choice of the Threshold Parameter in Wavelet Function Estimation. In A. Antoniadis & G. Oppenheim (Eds.), *Wavelets and Statistics* (pp. 261-299). NY: Springer-Verlag.

Nason, G. P. (1995b). The Stationary Wavelet Transform and Some Statistical Applications. In A. Antoniadis & G. Oppenheim (Eds.), *Wavelets and Statistics* (pp. 281-299). NY: Springer-Verlag.

Nason, G. P. (1996). Wavelet Shrinkage Using Cross-Validation. *Journal of the Royal Statistical Society, Series B, 58*, 463-479.

Nelson, P. R. C., Taylor, P. A. & MacGregor, J. F. (1996). Missing Data Methods in PCA and PLS: Score Calculations with Incomplete Observations. *Chemometrics and Intelligent Laboratory Systems, 35*, 45-65.

Nesic, Z., Davies, M. & Dumont, G. (1997). Paper Machine Data Analysis and Compression Using Wavelets. *Tappi Journal, 80*(10), 191-204.

Nikolaou, M. & Mantha, D. (1998). *Efficient Nonlinear Modeling Using Wavelets and Related Compression Techniques.* Paper presented at the NSF Workshop on Nonlinear Model Predictive Control, Ascona, CH.

Nikolaou, M. & Vuthandam, P. (1998). FIR Model Identification: Parsimony Through Kernel Compression with Wavelets. *AIChE Journal, 44*(1), 141-150.

Nikolaou, M. & You, Y. (1994). Use of Wavelets for Numerical Solution of Differential Equations. In R. L. Motard & B. Joseph (Eds.), *Wavelet Applications in Chemical Engineering* (pp. 209-274). Boston: Kluwer Academic Publishers.

Nomikos, P. & MacGregor, J. F. (1995). Multivariate SPC Charts for Monitoring Batch Processes. *Technometrics, 37*(1), 41-59.

Nordström, J.-E. P., Lindberg, S. & Lundström, A. (2002, 1-4, 2002). *Human Perception and Optical Measurement of HSWO Waviness.* Paper presented at the IPGAC (International Printing & Graphic Arts Conference), Bordeaux, France.

Nounou, M. N. & Bakshi, B. R. (1999). On-Line Multiscale Filtering of Random and Gross Errors Without Process Models. *AIChE Journal, 45*(5), 1041-1058.

Oppenheim, A. V., Schafer, R. W. & Buck, J. R. (1999). *Discrete-Time Signal Processing* (2nd ed.). Upper Saddle River (NJ): Prentice Hall.

Oussar, Y., Rivals, I., Personnaz, L. & Dreyfus, G. (1998). Training Wavelet Networks for Nonlinear Dynamic Input-Output Modeling. *Neurocomputing, 20*, 173-188.

Pal, S. K. & Mitra, M. (2004). *Pattern Recognition Algorithms for Data Mining*. Boca Raton: Chapman & Hall/CRC.

Pasti, L., Walczak, B., Massart, D. L. & Reschiglian, P. (1999). Optimization of Signal Denoising in Discrete Wavelet Transform. *Chemometrics and Intelligent Laboratory Systems, 48*, 21-34.

Pati, Y. C. & Krishnaprasad, P. S. (1993). Analysis and Synthesis of Feedforward Neural Networks Using Discrete Affine Wavelet Transformations. *IEEE Transactions on Neural Networks, 4*(1), 73-85.

Pati, Y. C., R. Rezaiifar, Krishnaprasad, P. S. & Dayawansa, W. P. (1993, Vov. 1-3, 1993). *A Fast Recursive Algorithm for System Identification and Model Reduction Using Rational Wavelets.* Paper presented at the Proceedings of the 27th Annual Asilomar Conference on Signals Systems and Computers.

Percival, D. B. & Walden, A. T. (2000). *Wavelets Methods for Time Series Analysis*. Cambridge (UK): Cambridge University Press.

Phatak, A. (1993). *Evaluation of Some Multivariate Methods and Their Applications in Chemical Engineering.* PhD Thesis, University of Waterloo.

Phatak, A., Reilly, P. M. & Penlidis, A. (1993). An Approach to Interval Estimation in Partial Least Squares Regression. *Analytica Chimica Acta, 277*, 495-501.

Pierna, J. A. F., Jin, L., Wahl, F., Faber, N. M. & Massart, D. L. (2003). Estimation of Partial Least Squares Regression Prediction Uncertainty When the Reference Values Carry a Sizeable Measurement Error. *Chemometrics and Intelligent Laboratory Systems, 65*, 281-291.

Plavajjhala, S., Motard, R. L. & Joseph, B. (1996). Process Identification Using Discrete Wavelet Transforms: Design of Prefilters. *AIChE Journal, 42*(3), 777-790.

Prosser, R. & Cant, R. S. (1998). On the Use of Wavelets in Computational Combustion. *Journal of Computational Physics, 147*, 337-361.

Qin, S. J. & Dunia, R. (2000). Determining the Number of Principal Components for Best Reconstruction. *Journal of Process Control, 10*, 245-250.

Ramanathan, J. & Zeitouni, O. (1991). On the Wavelet Transform of Fractional Brownian Motion. *IEEE Transactions on Information Theory, 37*(4), 1156-1158.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications* (2nd ed.). NY (etc.): Wiley.

Reis, M. S. (2000). *Introdução à Análise Multiresolução e suas Aplicações no Contexto da Engenharia Química.* Provas de Aptidão Pedagógica e Capacidade Científica, Universidade de Coimbra, Coimbra.

Reis, M. S. & Saraiva, P. M. (2003). *Practical Approaches for Conducing MSPC in Noisy Environments.* Paper presented at the PLS and Related Methods, Lisbon (Portugal).

Reis, M. S. & Saraiva, P. M. (2004a, 16-19 May). *Accounting for Measurement Uncertainties in Industrial Data Analysis.* Paper presented at the ESCAPE-14, European Symposium on Computer Aided Process Engineering, Lisbon (Portugal).

Reis, M. S. & Saraiva, P. M. (2004b). A Comparative Study of Linear Regression Methods in Noisy Environments. *Journal of Chemometrics, 18*(12), 526-536.

Reis, M. S. & Saraiva, P. M. (2004c). *A Comparative Study of the Performance of Various Linear Regression Methods in Noisy Environments.* Paper presented at the CAC-2004 Chemometrics in Analytical Chemistry, Lisbon (Portugal).

Reis, M. S. & Saraiva, P. M. (2005a). Heteroscedastic Latent Variable Modelling with Applications to Multivariate Statistical Process Control. *Chemometrics and Intelligent Laboratory Systems, In Press*.

Reis, M. S. & Saraiva, P. M. (2005b). *Integrating Data Uncertainty in Multiresolution Analysis.* Paper presented at the ESCAPE-15, European Symposium on Computer Aided Process Engineering, Barcelona (Spain).

Reis, M. S. & Saraiva, P. M. (2005c). Integration of Data Uncertainty in Linear Regression and Process Optimization. *AIChE Journal, 51*(11), 3007-3019.

Reis, M. S. & Saraiva, P. M. (2005d). *A Multiscale Approach for the Monitoring of Paper Surface Profiles.* Paper presented at the 5th Annual Meeting of ENBIS, Newcastle (UK).

Reis, M. S. & Saraiva, P. M. (2005e). Multiscale Statistical Process Control of Paper Surface Profiles. *Quality Technology & Quantitative Management, Accepted*.

Reis, M. S. & Saraiva, P. M. (2005f). *Multiscale Statistical Process Control of the Paper Surface.* Paper presented at the Gordon Research Conference on Statistics in Chemistry and Chemical Engineering, South Hadley, MA (USA).

Reis, M. S. & Saraiva, P. M. (2005g). *Multivariate and Multiscale Analysis of Paper Surface.* Paper presented at the CHEMPOR'2005, 9th International Chemical Engineering Conference.

Renaud, O., Starck, J.-L. & Murtagh, F. (2003). Prediction Based on a Multiscale Decomposition. *International Journal of Wavelets, Multiresolution and Information Processing, 1*(2), 217-232.

Renaud, O., Starck, J.-L. & Murtagh, F. (2005). Wavelet-Based Combined Signal Filtering and Prediction. *IEEE Transactions on Systems, Man, and Cibernetics (Part B: Cibernetics), in press, available at: http://www.unige.ch/~renaud/papers/abstractKalmanWavelet.html*.

Ricker, N. L. (1988). The Use of Biased Least-Squares Estimators for Parameters in Discrete-Time Pulse-Response Models. *Industrial & Engineering Chemistry Research, 27*, 343-350.

Río, F. J., Rio, J. & Rius, F. X. (2001). Prediction Intervals in Linear Regression Taking Into Account Errors in Both Axis. *Journal of Chemometrics, 15*, 773-788.

Rioul, O. & Vetterli, M. (1991). Wavelets and Signal Processing. *IEEE Signal Processing Magazine, 8*(4), 14-38.

Riu, J. & Rius, F. X. (1996). Assessing the Accuracy of Analytical Methods Using Linear Regression with Errors in both Axis. *Analytical Chemistry, 68*, 1851-1857.

Rooney, W. C. & Biegler, L. T. (2001). Design for Model Parameter Uncertainty Using Nonlinear Confidence Regions. *AIChE Journal, 47*(8), 1794-1804.

Rosen, C. (2001). *A Chemometric Approach to Process Monitoring and Control - With Applications to Wastewater Treatment Operation.* PhD Thesis, Lund University, Lund.

Rosen, C. & Lennox, J. A. (2001). Multivariate and Multiscale Monitoring of Wastewater Treatment Operation. *Water Research, 35*(14), 3402-3410.

Ruggeri, F. & Vidakovic, B. (1999). A Baysean Decision Theoretic Approach to the Choice of Thresholding Parameter. *Statistica Sinica, 9*, 183-197.

Safavi, A. A., Chen, J. & Romagnoli, J. A. (1997). Wavelet-Based Density Estimation and Application to Process Monitoring. *AIChE Journal, 43*(5), 1227-1241.

Safavi, A. A. & Romagnoli, J. A. (1997). Application of Wavelet-Based Neural Networks to the Modelling and Optimization of an Experimental Distillation Column. *Engineering Applications of Artificial Intelligence, 10*(3), 301-313.

Sander, M. (1991). *A Practical Guide to the Assessment of Surface Texture*. Göttingen: Feinprüf Perthen GmbH.

Santos, J. C., Cruz, P., Magalhães, F. D. & Mendes, A. (2003). 2-D Wavelet-Based Adaptive-Grid Method for the Resolution of PDEs. *AIChE Journal, 49*(3), 706-717.

Saraiva, P. M. (1993). *Data-Driven Learning Frameworks for Continuous Process Analysis and Improvement.* PhD Thesis, Massachusetts Institute of Technology.

Saraiva, P. M. & Stephanopoulos, G. (1992). Continuous Process Improvement through Inductive and Analogical Learning. *AIChE Journal, 38*(2), 161-183.

Saraiva, P. M. & Stephanopoulos, G. (1998). Process Improvement: an Exploratory Data Analysis Approach within an Interval-Based Optimization Framework. *Production and Operations Management (POMS), 7*(1), 19-37.

Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.

Seber, G. A. F. & Wild, C. J. (1989). *Nonlinear Regression*. New York: Wiley.

Shao, R., Jia, F., Martin, E. B. & Morris, A. J. (1999). Wavelets and Non-Linear Principal Components Analysis for Process Monitoring. *Control Engineering Practice, 7*, 865-879.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product* (Vol. Republished in 1980 as a 50th Anniversary Commemorative Reissue by ASQC Quality Press). New York: D. Van Nostrand Company, Inc.

Shi, R. & MacGregor, J. F. (2000). Modeling of Dynamic Systems Using Latent Variable and Subspace Methods. *Journal of Chemometrics, 14*, 423-439.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P.-Y., et al. (1995). Nonlinear Black-box Modeling in System Identification: a Unified Overview. *Automatica, 31*(12), 1691-1724.

Smilde, A. K., Westerhuis, J. A. & de Jong, S. (2003). A Framework for Sequential Multiblock Component Methods. *Journal of Chemometrics, 17*, 323-337.

Soares, M. J. F. E. (1997). *Introdução à Teoria das Onduletas*: Universidade do Minho.

Starck, J.-L., Murtagh, F. & Bijaoui, A. (1998). *Image Processing and Data Analysis - The Multiscale Approach*. Cambridge: Cambridge University Press.

Staszewski, W. J. (1998). Wavelet Based Compression and Feature Selection for Vibration Analysis. *Journal of Sound and Vibration, 211*(5), 735-760.

Stephanopoulos, G., Dyer, M. & Karsligil, O. (1997a). Multi-Scale Modeling, Estimation and Control of Processing Systems. *Computers and Chemical Engineering, 21*(Supplement for the PSE'97-ESCAPE-7, joint 6th International Symposium of Process Systems Engineering and 30th European Symposium on

Computer Aided Process Engineering, May 1997, Trondheim, Norway), S797-S803.

Stephanopoulos, G., Karsligil, O. & Dyer, M. (1997b). *A Multi-Scale Systems Theory for Process Estimation and Control* (Basis report of a paper presented at the NATO-ASI on "Nonlinear Model Based Process Control", 10-20 August 1997, Antalya, Turkey).

Stephanopoulos, G., Karsligil, O. & Dyer, M. (2000). Multi-Scale Aspects in Model-Predictive Control. *Journal of Process Control, 10*, 275-282.

Strang, G. & Nguyen, T. (1997). *Wavelets and Filter Banks*. Wellesley (MA): Wellesley-Cambridge Press.

Sun, W., Palazoğlu, A. & Romagnoli, J. A. (2003). Detecting Abnormal Process Trends by Wavelet-Domain Hidden Markov Models. *AIChE Journal, 49*(1), 140-150.

Sze, D. H. & Waech, T. G. (1997). *Paper Waviness and Finer Alignment Streaks in Paper Forming.* Paper presented at the CPPA Annual Technical Meeting Proceedings, Montréal, Canada.

Tangirala, A. K. (2001). *Multirate Control and Multiscale Monitoring of Chemical Processes.* PhD Thesis, University of Alberta.

Teppola, P. & Minkkinen, P. (2000). Wavelet-PLS Regression Models for Both Exploratory Data Analysis and Process Monitoring. *Journal of Chemometrics, 14*, 383-399.

Teppola, P. & Minkkinen, P. (2001). Wavelets for Scrutinizing Multivariate Exploratory Models - Interpreting Models Through Multiresolution Analysis. *Journal of Chemometrics, 15*, 1-18.

Tewfik, A. H. (1992). Correlation Structure of the Discrete Wavelet Coefficients of Fractional Brownian Motion. *IEEE Transactions on Information Theory, 38*(2), 904-909.

Tewfik, A. H., Sinha, D. & Jorgensen, P. (1992). On the Optimal Choise of a Wavelet for Signal Representation. *IEEE Transactions on Information Theory, 38*(2), 747-765.

Theodoridis, S. & Koutroumbas, K. (2003). *Pattern Recognition* (2nd ed.). Amsterdam: Elsevier.

Thomas, E. V. (2003). Non-Parametric Statistical Methods for Multivariate Calibration Model Selection and Comparison. *Journal of Chemometrics, 17*, 653-659.

Top, S. & Bakshi, B. R. (1998). *Improved Statistical Process Control Using Wavelets.* Paper presented at the Foundation of Computer Aided Process Operations FOCAPO 98.

Tracy, N. D., Young, J. C. & Mason, R. L. (1992). Multivariate Control Charts for Individual Observations. *Journal of Quality Technology, 24*(2), 88-95.

Trygg, J. (2001). *Parsimonious Multivariate Models.* PhD Thesis, Umeå University.

Trygg, J., Kettaneh-Wold, N. & Wallbäcks, L. (2001). 2D Wavelet Analysis and Compression of On-Line Industrial Process Data. *Journal of Chemometrics, 15*, 299-319.

Trygg, J. & Wold, S. (1998). PLS Regression on Wavelet Compressed NIR Spectra. *Chemometrics and Intelligent Laboratory Systems, 42*, 209-220.

Tsatsanis, M. K. & Giannakis, G. B. (1993). Time-Varying System Identification and Model Validation Using Wavelets. *IEEE Transactions on Signal Processing, 41*(12), 3512-3523.

Tsuge, Y., Hiratsuka, K., Takeda, K. & Matsuyama, H. (2000). A Fault Detection and Diagnosis for the Continuous Process with Load-Fluctuations Using Orthogonal Wavelets. *Computers and Chemical Engineering, 24*, 761-767.

Ungarala, S. & Bakshi, B. R. (2000). A Multiscale, Baysean and Error-In-Variables Approach fot Linear Dynamic Data Rectification. *Computers and Chemical Engineering, 24*, 445-451.

Valle, S., Li, W. & Qin, S. J. (1999). Selection of the Number of Principal Components: The Variance of the Reconstruction Error Criterion with a Comparison to Other Methods. *Industrial & Engineering Chemistry Research, 38*, 4389-4401.

Van Eperen, R. H. (1991). *Paper Properties and Testing Procedures* (3rd ed. Vol. 7). Atlanta: The Joint Textbook Committee of The Paper Industry.

Vannucci, M., Sha, N. & Brown, P. J. (2005). NIR and Mass Spectra Classification: Baysean Methods for Wavelet-Based Feature Selection. *Chemometrics and Intelligent Laboratory Systems, 77*, 139-148.

Veitch, D. & Abry, P. (1999). A Wavelet-Based Joint Estimator of the Parameters of Long-Range Dependence. *IEEE Transactions on Information Theory, 45*(3), 878-897.

Vetterli, M. & Kovačević, J. (1995). *Wavelets and Subband Coding*. New Jersey: Prentice Hall.

Vidakovic, B. & Ruggeri, F. (2001). BAMS Method: Theory and Simulations. *Sankhyā: The Indian Journal of Statistics, 63*(Part 2), 234-249.

Vogt, F. & Mizaikoff, B. (2003). Dynamic Determination of the Dimension of PCA Calibration Models Using F-Statistics. *Journal of Chemometrics, 17*, 346-357.

Vogt, F. & Tacke, M. (2001). Fast Principal Component Analysis of Large Data Sets. *Chemometrics and Intelligent Laboratory Systems, 59*, 1-18.

Wågberg, P. & Johansson, P.-Å. (2002). Characterization of Paper Surfaces Using Optical Profilometry. In J. Borch, M. B. Lyne, R. E. Mark & C. C. Habeger, Jr. (Eds.), *Handbook of Physical Testing of Paper* (2nd ed., Vol. 2). New York: Marcel Dekker.

Walczak, B. & Massart, D. L. (1997a). Noise Suppression and Signal Compression Using the Wavelet Packet Transform. *Chemometrics and Intelligent Laboratory Systems, 36*, 81-94.

Walczak, B. & Massart, D. L. (1997b). Wavelet Packet Transform Applied to a Set of Signals: A New Approach to the Best-Basis Selection. *Chemometrics and Intelligent Laboratory Systems, 38*, 39-50.

Walczak, B. & Massart, D. L. (2001). Dealing with Missing Data. *Chemometrics and Intelligent Laboratory Systems, 58*, Part I: 15-27, Part II: 29-42.

Walczak, B., van den Bogaert, V. & Massart, D. L. (1996). Application of Wavelet Packet Transform in Pattern Recognition of Near-IR Data. *Analytical Chemistry, 68*, 1742-1747.

Walker, J. S. (1999). *A Primer on Wavelets and their Scientific Applications*. Boca Raton: Chapman & Hall /CRC.

Walter, G. G. (1994). *Wavelets and Other Orthogonal Systems With Applications*. Boca Raton: CRC Press.

Wang, X. Z. (1999). *Data Mining and Knowledge Discovery for Process Monitoring and Control*. London: Springer.

Wang, X. Z., Chen, B. H., Yang, S. H. & McGreavy, C. (1999). Application of Wavelets and Neural Networks to Diagnostic System Development, 1, Feature Extraction. *Computers and Chemical Engineering, 23*, 899-906.

Watson, G. H., Gilholm, K. & Jones, J. G. (1999). A Wavelet-Based Method for Finding Inputs of Given Energy which Maximize the Outputs of Nonlinear Systems. *International Journal of Systems Science, 30*(12), 1297-1307.

Wentzell, P. D., Andrews, D. T., Hamilton, D. C., Faber, K. & Kowalski, B. R. (1997a). Maximum Likelihood Principal Component Analysis. *Journal of Chemometrics, 11*, 339-366.

Wentzell, P. D., Andrews, D. T. & Kowalski, B. R. (1997b). Maximum Likelihood Multivariate Calibration. *Analytical Chemistry, 69*, 2299-2311.

Wentzell, P. D. & Lohnes, M. T. (1999). Maximum Likelihood Principal Component Analysis with Correlated Measurements Errors: Theoretical and Practical Considerations. *Chemometrics and Intelligent Laboratory Systems, 45*, 65-85.

Westerhuis, J. A., Gurden, S. P. & Smilde, A. K. (2000). Generalized Contribution Plots in Multivariate Statistical Process Monitoring. *Chemometrics and Intelligent Laboratory Systems, 51*, 95-114.

Westerhuis, J. A., Kourti, T. & MacGregor, J. F. (1998). Analysis of Multiblock and Hierarchical PCA and PLS Models. *Journal of Chemometrics, 12*, 301-321.

Westerhuis, J. A., Kourti, T. & MacGregor, J. F. (1999). Comparing Alternative Approaches for Multivariate Statistical Analysis of Batch Process Data. *Journal of Chemometrics, 13*, 397-413.

Wharton, J. A., Wood, R. J. K. & Mellor, B. G. (2003). Wavelet Analysis of Electrochemical Noise Measurements During Corrosion of Austenitic and Superduplex Stainless Steels in Chloride Media. *Corrosion Science, 45*, 97-122.

Whitcher, B. (2004). Wavelet-Based Estimation for Seasonal Long-Memory Processes. *Technometrics, 46*(2), 225-238.

Whitcher, B. J. (1998). *Assessing Nonstationary Time Series Using Wavelets.* PhD Thesis, University of Washington.

Willsky, A. S. (2002). Multiresolution Markov Models for Signal and Image Processing. *Proceedings of the IEEE, 90*(8), 1396-1458.

Wise, B. W. & Gallagher, N. B. (1996). The Process Chemometrics Approach to Process Monitoring and Fault Detection. *Journal of Process Control, 6*(6), 329-348.

Wold, S. (1978). Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics, 20*(4), 397-405.

Wold, S., Berglung, A. & Kettaneh, N. (2002). New and Old Trends in Chemometrics. How to Deal With the Increasing Data Volumes in R&D&P (Research Development and Production) – With Examples from Pharmaceutical Research and Process Modelling. *Journal of Chemometrics, 16*, 377-386.

Wold, S., Sjöström, M., Carlson, R., Lundstedt, T., Hellberg, S., Skagerberg, B., et al. (1986). Multivariate Design. *Analytica Chimica Acta, 191*, 17-32.

Wold, S., Sjöström, M. & Eriksson, L. (2001). PLS-Regression: A Basic Tool of Chemometrics. *Chemometrics and Intelligent Laboratory Systems, 58*, 109-130.

Woodall, W. H., Spitzner, D. J., Montgomery, D. C. & Gupta, S. (2004). Using Control Charts to Monitor Process and Product Quality Profiles. *Journal of Quality Technology, 36*(3), 309-320.

Wornell, G. W. (1990). A Karhunen-Loève-like Expansion for 1/f Processes via Wavelets. *IEEE Transactions on Information Theory, 36*(4), 859-861.

Wornell, G. W. & Oppenheim, A. V. (1992). Estimation of Fractal Signals from Noisy Measurements Using Wavelets. *IEEE Transactions on Signal Processing, 40*(3), 611-623.

Yacoub, F. & MacGregor, J. F. (2004). Product Optimization and Control in the Latent Variable Space of Nonlinear PLS Models. *Chemometrics and Intelligent Laboratory Systems, 70*, 63-74.

Yoon, S. & MacGregor, J. F. (2001, 4-6 June). *Unifying PCA and Multiscale Approaches to Fault Detection and Isolation.* Paper presented at the 6th IFAC Symposium on Dynamics and Control of Process Systems, Korea.

Yoon, S. & MacGregor, J. F. (2004). Principal-Component Analysis of Multiscale Data for Process Monitoring and Fault Diagnosis. *AIChE Journal, 50*(11), 2891-2903.

Yu, H. & MacGregor, J. F. (2004a). *Digital Imaging for Process Monitoring and Control with Industrial Applications.* Paper presented at the 7th International Symposium on Advanced Control of Chemical Processes - ADCHEM.

Yu, H. & MacGregor, J. F. (2004b). Monitoring Flames in an Industrial Boiler Using Multivariate Image Analysis. *AIChE Journal, 50*(7), 1474-1483.

Zhang, Q. (1995). Wavelets and Regression Analysis. In A. Antoniadis & G. Oppenheim (Eds.), *Wavelets and Statistics* (pp. 399-407). NY.

Zhang, Q. & Benveniste, A. (1992). Wavelet Networks. *IEEE Transactions on Neural Networks, 3*(6), 889-898.

Zhao, J., Chen, B. & Schen, J. (1998). Multidimensional Non-Orthogonal Wavelet-Sigmoid Basis Function Neural Network for Dynamic Process Fault Diagnosis. *Computers and Chemical Engineering, 23*, 83-92.

Zhao, Z., Jin, Y. & Wang, J. (2000). Application of Wavelet Transform to Process Operating Region Recognition. *Journal of Chemical Engineering of Japan, 33*(6), 823-831.

Zhong, H., Zhang, J., Gao, M., Zheng, J., Li, G. & Chen, L. (2001). The Discrete Wavelet Neural Network and its Application in Oscillographic Chronopotentiometric Determination. *Chemometrics and Intelligent Laboratory Systems, 59*, 67-74.

Appendices

Appendix A. Additional Information Regarding MLMLS Method

In this appendix, we provide further information regarding the motivation underlying the development of the MLMLS method, and present some results that illustrate its relationship with other related methods and its potential. For the sake of simplifying the exposition, without compromising rigour and generality, only the univariate case will be addressed here.

A.1 EIV Formulation of the Linear Regression Problem

The classical EIV model consists of the following functional relationship linking the "true" values of the predictor (η_i) and response (ξ_i) variables,

$$\eta_i = \beta_0 + \beta_1 \xi_i \tag{A.1}$$

Along with the measurement equations,

$$\begin{aligned} x_i &= \xi_i + \delta_i \\ y_i &= \eta_i + \varepsilon_i \end{aligned} \tag{A.2}$$

Let us assume that $\delta_i \sim N\left(0,\sigma_\delta^2\right), \varepsilon_i \sim N\left(0,\sigma_\varepsilon^2\right)$, with δ_i and ε_i statistically independent. Inserting (A.2) into (A.1), and rearranging terms, the following relationship can be obtained, linking the measured quantities:

$$y_i = \beta_0 + \beta_1 x_i + \left(\varepsilon_i - \beta_1 \delta_i\right) \tag{A.3}$$

As the error term, $\varepsilon_i - \beta_1 \delta_i$, is not independent of the quantity $y_i = \beta_0 + \beta_1 x_i$, one can not estimate the model parameters using classical least squares (e.g. Draper & Smith, 1998, p. 90).

A.2 The Berkson Case (Controlled Regressors with Error)

Berkson pointed out that in many experiments the above referred correlation does not exist because there are circumstances where x_i is a "controlled quantity", i.e., a set-point or target for the predictor variable that we would like to keep fixed during the realization of the trial, but, due to experimental limitations, we can not achieve such a goal. Thus, the "observed" value of the predictor variable, x_i, is directly controlled by the experimenter, while the true values, ξ_i, are unknown and may experiment some variation. In this situation, it can be assumed that the "true" predictor is scattered around the target value as follows:

$$\xi_i = x_i + \delta_i \tag{A.4}$$

The measured value of the response is subject to random error, according to:

$$y_i = \eta_i + \varepsilon_i \tag{A.5}$$

Once again, we assume $\delta_i \sim N\left(0,\sigma_\delta^2\right), \varepsilon_i \sim N\left(0,\sigma_\varepsilon^2\right)$, with δ_i and ε_i statistically independent. The true value of the response still depends upon the true value of the predictor variable, according to the functional relationship: $\eta_i = \beta_0 + \beta_1 \xi_i$. From (A.4), (A.5) and this functional relationship between the true values, the following relationship between the "measured" quantities can be derived:

$$y_i = \beta_0 + \beta_1 x_i + \left(\varepsilon_i + \beta_1 \delta_i\right) \tag{A.6}$$

where the error term, $\varepsilon_i + \beta_1 \delta_i$, is now independent of the quantity $y_i = \beta_0 + \beta_1 x_i$. Berkson argues that the requirements of the classical least squares case are now fulfilled, and we can use it to estimate the model parameters β_0 and β_1. The optimization formulation proposed in Section 2.1.1 results from deriving the log-likelihood function for the above situation in the heteroscedastic case, under the assumption of Gaussian errors.

A.3 Results

The following results illustrate relationships of the MLMLS approach with related methods, such as OLS and MLS. Two versions of MLS are tested, designated as MLS and MLS (Rovira), to check the correctness of the first one (written by the author in Matlab code) with a version independently developed by the research group from Chemometrics and Qualimetrics centre, at the Universitat Rovira i Virgili (Spain, available at http://www.quimica.urv.es/quimio/ang/maincat.html).

We consider a model following the Berkson assumptions with parameters $b_0 = 2$ and $b_1 = 4$. The model parameters are repeatedly estimated from 100 observations, and the errors obtained in 500 of such realizations are presented in the following figures.

A.3.1 No errors in X, homoscedastic errors in Y

Let us first consider the situation where the controlled regressors are not affected by any sort of error and the response is affected by homoscedastic errors with variance

$\sigma_Y^2 = 0.16$. In these circumstances, the Berkson model essentially reduces to the classical OLS model, and therefore all the methods should have similar performances, as all of them can handle this model, at least as a particular situation, and there are no numerical issues to be considered in the univariate case, as can be seen in Figure A-1.

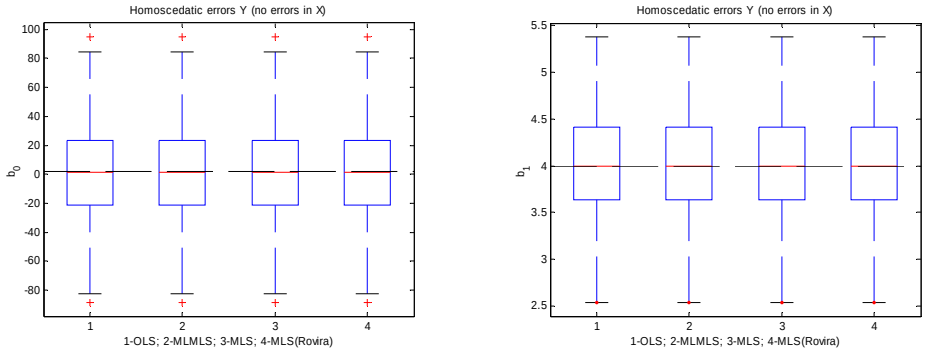


Figure A-1. Parameter estimates for the case: "no errors in X, homoscedastic errors in Y". The true values for the parameters in the Berkson model are indicated by horizontal lines.

A.3.2 Homoscedastic errors in X and Y

Considering the situation where both the controlled regressors and response are affected by homoscedastic errors with variances $\sigma_X^2 = \sigma_Y^2 = 0.16$ and $\sigma_Y^2 = 0.16$ (Figure A-2), it is possible to verify that the MLS method provides biased estimates and with higher variance relatively to other methods, something that can be attributed to the mismatch between the model structure assumed by this method for the data generating process and that actually underlying analysed data. OLS is more robust in this regard, in spite of presenting a slight tendency towards an increased variance in the estimates.

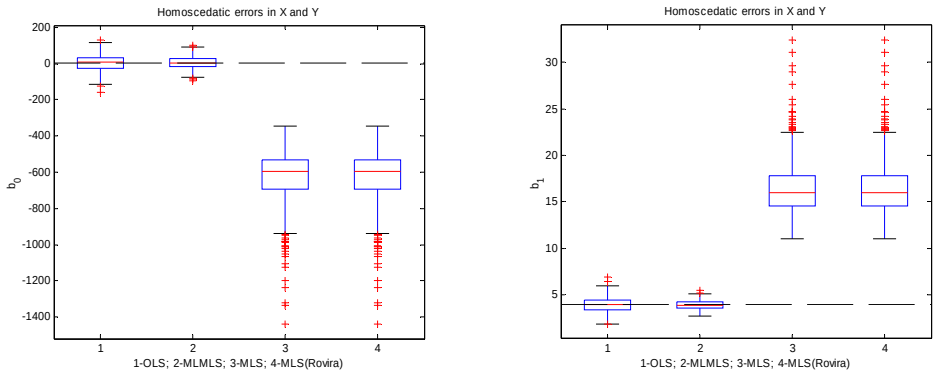


Figure A-2. Parameter estimates for the case: "Homoscedastic errors in X and Y".

A.3.3 Heteroscedastic errors in X and Y

Considering finally the situation where both the controlled regressors are affected by heteroscedastic errors of the proportional type (Figure A-3), it is possible to see that the MLMLS is now clearly the best method among all the alternatives tested (unbiased parameter estimates with lower associated variances).

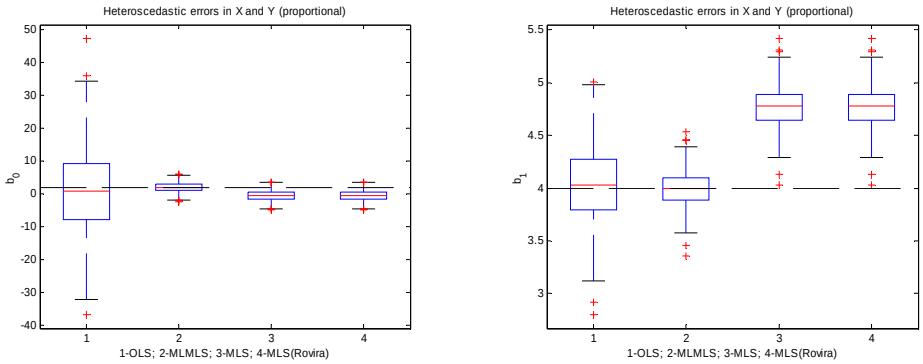


Figure A-3. Parameters estimates for the case: "Heteroscedastic errors in X and Y" (proportional type).

Therefore, we can conclude that under the scope of a data generating process following Berkson's assumptions, the MLMLS method always leads to unbiased estimates with variance that is at least as low as that for any other of the tested methods, i.e., never performs worse than its counterparts, and in fact can perform significantly better under more complex errors structures, thus attesting the suitability of this estimation procedure, based on the maximization of the log-likelihood function.

Appendix B. Analytical Derivation for the Gradients of Λ

In this section we present a derivation for the gradients of the log-likelihood function, Λ, in order to the parameter vectors to be estimated under the maximum likelihood approach, namely:[45]

- $\underline{\mu}_X$, the mean vector;
- $\underline{\lambda} = diag(\Delta_l^{*1/2})$, i.e., the vector of diagonal elements of $\Delta_l^{*1/2}$ ($\Delta_l^{*1/2}$ is a diagonal matrix such that $\Delta_l = \Delta_l^{*1/2} \cdot \Delta_l^{*1/2}$);
- $\underline{\alpha}$, the vector of rotation angles to be applied to the initial estimate of A (A_0).

Let us first clarify the conventions to be followed during the course of derivations. The first convention regards the definition of the derivative of a matrix F(X), $m \times p$, in order to another matrix, X, $n \times q$ (Magnus & Neudecker, 1988):

[45] The single underscore used below some of the above quantities has the purpose of highlighting their vector nature, to avoid any confusion with scalar quantities with similar notations.

$$D_X F(X) = \frac{\partial\, vec\, F(X)}{\partial (vec\, X)^T} \tag{B.1}$$

where *vec* is the operator that vectorizes a matrix, by successively stacking its columns, one below the others, starting from the first column.

Thus, $D_X F(X)$ is a $mp \times nq$ matrix, whose element (*i,j*) is the partial derivative of the function at the i^{th}-entry of $vec\, F(X)$ in order to the variable at the j^{th} entry of $vec\, X$.

Another useful definition regards the extension of the notion of differential to matrix quantities:

$$d\, vec\, F(X) = A(X)\, d\, vec\, X \tag{B.2}$$

According to the *identification theorem* for matrix functions (Magnus & Neudecker, 1988), the existence of (B.2) implies and is implied by

$$D_X F(X) = A(X) \tag{B.3}$$

Thus, the definition of differential and the identification theorem taken together are instrumental in the calculation of derivatives. The basic procedure here adopted, following Magnus & Neudecker (1988) guidelines, is as follows: *i.* compute the differential; *ii.* vectorize the expression obtained in *i.*; *iii.* use the identification theorem to obtain the derivative.

B.1 Derivation of the Gradients

For the sake of clarity, the elements or building blocks appearing in the expression of the log-likelihood will be isolated. When rewritten in terms of the quantities to be estimated by the numerical algorithm, they have the following form:

$$\Lambda(\mu_X, \underline{\lambda}, \underline{\alpha}) = C - \frac{1}{2}\sum_{k=1}^{n_{obs}} \ln\left|\Sigma_x(k)\right| - \frac{1}{2}\sum_{k=1}^{n_{obs}}\left[(x(k)-\mu_X)^T \Sigma_x^{-1}(k)(x(k)-\mu_X)\right]$$

where, (B.4)

$$\begin{cases} \Sigma_x(k) = A\Delta_l A^T + \Delta_m(k) \\ A = R(\underline{\alpha})A_0 \\ \Delta_l = \Delta_l^{*1/2} \cdot \Delta_l^{*1/2} \\ \Delta_l^{*1/2} = diag(\underline{\lambda}) \end{cases}$$

with C being a constant, *diag* the operator that when applied to a square matrix produces a vector with its diagonal elements and that, when applied to a vector, produces a square diagonal matrix with the elements of the vector along the main diagonal.

The basic elements identified in (B.4) are:

- $\ln\left|\Sigma_x(k)\right|$;
- $(x(k)-\mu_X)^T \Sigma_x^{-1}(k)(x(k)-\mu_X)$.

The derivation of the expression for their gradients, in order to the parameter vectors, is systematized in the following steps:

- *1.i*. Derivation of differential and gradients for $\Sigma_x(k)$;
- *1.ii*. Derivation of differential and gradients for $\ln\left|\Sigma_x(k)\right|$;
- *2.i*. Derivation of differential and gradients for $\Sigma_x^{-1}(k)$;
- *2.ii*. Derivation of differential and gradients for $(x(k)-\mu_X)^T \Sigma_x^{-1}(k)(x(k)-\mu_X)$;
- *3.* Derivation of the gradients for Λ.

B.1.1 Derivation of the differential and gradients for $\Sigma_x(k)$ (1.i)

As the sums do not appear in the building blocks, we will drop the sum index, *k*, keeping in mind that these quantities vary with the observation index.

From (B.4),

$$d\,\Sigma_x = d\left(A\Delta_l A^T\right) + d\left(\Delta_m\right) = d\left(A\Delta_l A^T\right) \tag{B.5}$$

($d\left(\Delta_m\right) = \underline{\underline{0}}$ because it is a constant matrix). Defining $A^* = A\Delta^{*1/2}$, we get from Magnus & Neudecker (1988, p.182):

$$d\,vec\,\Sigma_x = d\,vec\left(A^* A^{*T}\right) = 2N_n(A^* \otimes I_n)\,d\,vec\,A^* \tag{B.6}$$

where $N_n = \frac{1}{2}\left(I_{n^2} + K_{n,n}\right)$ ($K_{n,n}$ the commutation matrix), I_n is the identity matrix with the dimension given by the subscript (n in this case) and $\otimes$ is the Kronecker product.

Now, as $A^* = A\Delta^{*1/2}$,

$$d\,A^* = d(A\Delta^{1/2*}) = d(A)\Delta^{1/2*} + Ad(\Delta^{1/2*}) \tag{B.7}$$

Vectorzing (B.7),

$$\begin{aligned} d\,vec\,A^* &= vec\left\{d(A)\Delta^{1/2*}\right\} + vec\left\{Ad(\Delta^{1/2*})\right\} \\ \Leftrightarrow d\,vec\,A^* &= \left(\Delta^{1/2*}\right)^T \otimes I_n\,d\,vec\,A + \left(I_p \otimes A\right)d\,vec\,\Delta^{1/2*} \end{aligned} \tag{B.8}$$

(cf. Magnus & Neudecker 1988, p.31).

We see that, introducing (B.8) in (B.6), it is possible to express $d\,vec\,\Sigma_x$ in terms of $d\,vec\,A$ and $d\,vec\,\Delta^{1/2*}$. Let us now see how we can express these quantities in terms of the parameter vectors $\underline{\mu}_X$, $\underline{\lambda}$, and $\underline{\alpha}$, beginning with $d\,vec\,A$. As $A = R(\underline{\alpha})A_0$,

$$\begin{aligned} & dA = d\left\{ R(\underline{\alpha})\right\} A_0 \\ & \Rightarrow d\,vec\,A = \left(A_0^{\,T} \otimes I_n \right) d\,vec\,R(\underline{\alpha}) \end{aligned} \tag{B.9}$$

Using Wentzell *et al.* (1997a) result for $d\,vec\,R(\underline{\alpha})$, and defining matrix $\tilde{\mathbf{G}}$ as:[46]

$$\tilde{G} = \left[vec\,G_1 R(\underline{\alpha}) \quad vec\,G_2 R(\underline{\alpha}) \cdots vec\,G_{n-1} R(\underline{\alpha}) \right] \tag{B.10}$$

we find that,

$$d\,vec\,R(\underline{\alpha}) = \tilde{\mathbf{G}}\, d\underline{\alpha} \tag{B.11}$$

On the other hand, $\Delta^{1/2*} = diag(\underline{\lambda})$, and it is possible to prove that,

$$d\,vec\,\Delta^{1/2*} = \mathbf{T} d\underline{\lambda} \tag{B.12}$$

where $\mathbf{T}$ is a matrix with a sparse structure, such that $vec\,\Delta^{1/2*} = \mathbf{T}\underline{\lambda}$.

Introducing (B.11) in (B.9), and entering with the result of this substitution, together with (B.12), in (B.8), we can now specify entirely the differential $d\,vec\,\Sigma_x$ presented in (B.6):

$$\begin{aligned} d\,vec\,\Sigma_x = {} & 2N_n(A^* \otimes I_n)\left(\Delta^{1/2*}\right)^T \otimes I_n \left(A_0^{\,T} \otimes I_n \right) \tilde{\mathbf{G}}\, d\underline{\alpha} + \\ & + 2N_n(A^* \otimes I_n)\left(I_p \otimes A \right) \mathbf{T} d\underline{\lambda} \end{aligned} \tag{B.13}$$

---

[46] See Wentzell *et al.* (1997a), p.365 for a definition of matrices $\{G_i\}_{i=1,n-1}$.

From the identification theorem for matrix functions, we can then calculate the gradients,

$$D_{\underline{\alpha}}\Sigma_x = 2N_n(A^* \otimes I_n)\left(\Delta^{1/2*}\right)^T \otimes I_n \left(A_0{}^T \otimes I_n\right)\tilde{\mathbf{G}} \tag{B.14}$$

$$D_{\underline{\lambda}}\Sigma_x = 2N_n(A^* \otimes I_n)\left(I_p \otimes A\right)\mathbf{T} \tag{B.15}$$

B.1.2 Derivation of the differential and gradients for $\ln|\Sigma_x(k)|$ (1.ii)

The differential of the natural logarithm of (a positive) scalar variable ϕ is,

$$d\ln\phi = \frac{1}{\phi}d\phi \tag{B.16}$$

In the present case, $\phi = |\Sigma_x|$, which means that,

$$d\ln|\Sigma_x| = \frac{1}{|\Sigma_x|}d|\Sigma_x| \tag{B.17}$$

Now, in Magnus & Neudecker (1988, p.178) we can find the expression for $D_X|X|$, that leads to the following differential for $|\Sigma_x|$:

$$d|\Sigma_x| = |\Sigma_x|\left\{vec\left(\Sigma_x{}^{-1}\right)^T\right\}^T d\,vec\,\Sigma_x \tag{B.18}$$

Thus, substituting (B.18) in (B.17),

$$d \ln\left|\Sigma_x\right| = \left\{ vec\left(\Sigma_x^{-1}\right)^T \right\}^T d\, vec\, \Sigma_x \tag{B.19}$$

This also leads us to the result that for $D_{\Sigma_x} \ln\left|\Sigma_x\right|$ (using the identification theorem):

$$D_{\Sigma_x} \ln\left|\Sigma_x\right| = \left\{ vec\left(\Sigma_x^{-1}\right)^T \right\}^T \tag{B.20}$$

Finally, introducing (B.13) in (B.19), we get the fully expanded expression for the differential of $\ln\left|\Sigma_x\right|$, in terms of the parameters vectors,

$$\begin{aligned} d \ln\left|\Sigma_x\right| = & \left\{ vec\left(\Sigma_x^{-1}\right)^T \right\}^T 2N_n(A^* \otimes I_n)\left(\Delta^{1/2*}\right)^T \otimes I_n \left(A_0^{\,T} \otimes I_n\right)\tilde{\mathbf{G}}\, d\underline{\alpha} + \\ & + \left\{ vec\left(\Sigma_x^{-1}\right)^T \right\}^T 2N_n(A^* \otimes I_n)\left(I_p \otimes A\right)\mathbf{T} d\underline{\lambda} \end{aligned} \tag{B.21}$$

which means that the respective gradients are:

$$\begin{aligned} D_{\underline{\alpha}} \ln\left|\Sigma_x\right| &= 2\left\{ vec\left(\Sigma_x^{-1}\right)^T \right\}^T N_n(A^* \otimes I_n)\left(\Delta^{1/2*}\right)^T \otimes I_n \left(A_0^{\,T} \otimes I_n\right)\tilde{\mathbf{G}} \\ D_{\underline{\lambda}} \ln\left|\Sigma_x\right| &= 2\left\{ vec\left(\Sigma_x^{-1}\right)^T \right\}^T N_n(A^* \otimes I_n)\left(I_p \otimes A\right)\mathbf{T} \end{aligned} \tag{B.22}$$

B.1.3 Derivation of differential and gradients for $\Sigma_x^{-1}(k)$ (2.i)

From Magnus & Neudecker (1988, p.183):

$$d\, vec\, \Sigma_x^{-1} = -\left[\left(\Sigma_x^{\,T}\right)^{-1} \otimes \Sigma_x^{-1} \right] d\, vec\, \Sigma_x \tag{B.23}$$

Therefore, from (B.13):

$$\begin{aligned} d\,vec\,\Sigma_x^{-1} = &-2\left[\left(\Sigma_x^{T}\right)^{-1}\otimes\Sigma_x^{-1}\right]N_n(A^*\otimes I_n)\left(\Delta^{1/2*}\right)^T\otimes I_n\left(A_0^{T}\otimes I_n\right)\tilde{\mathbf{G}}\,d\underline{\alpha}+ \\ &-2\left[\left(\Sigma_x^{T}\right)^{-1}\otimes\Sigma_x^{-1}\right]N_n(A^*\otimes I_n)\left(I_p\otimes A\right)\mathbf{T}d\underline{\lambda} \end{aligned} \tag{B.24}$$

and,

$$\begin{aligned} D_{\underline{\alpha}}\,vec\,\Sigma_x^{-1} &= -2\left[\left(\Sigma_x^{T}\right)^{-1}\otimes\Sigma_x^{-1}\right]N_n(A^*\otimes I_n)\left(\Delta^{1/2*}\right)^T\otimes I_n\left(A_0^{T}\otimes I_n\right)\tilde{\mathbf{G}} \\ D_{\underline{\lambda}}\,vec\,\Sigma_x^{-1} &= -2\left[\left(\Sigma_x^{T}\right)^{-1}\otimes\Sigma_x^{-1}\right]N_n(A^*\otimes I_n)\left(I_p\otimes A\right)\mathbf{T} \end{aligned} \tag{B.25}$$

B.1.4 Derivation of differential and gradients for $(x(k)-\underline{\mu}_X)^T\,\Sigma_x^{-1}(k)(x(k)-\underline{\mu}_X)$ (2.ii)

Writing $(x(k)-\underline{\mu}_X)^T\,\Sigma_x^{-1}(k)(x(k)-\underline{\mu}_X)$ as $\Delta x^T\Sigma_x^{-1}\Delta x$, we can use the well known result from derivative of a quadratic expression, x^TAx, in order to x, to find the term of the differential regarding Δx (cf. Magnus & Neudecker 1988, p.177), as well as the result already derived for Σ_x^{-1}, for the remaining term,[47]

$$d\left(\Delta x^T\Sigma_x^{-1}\Delta x\right) = \Delta x^T(\Sigma_x^{-1}+\left(\Sigma_x^{-1}\right)^T)\,d\Delta x+\Delta x^T\otimes\Delta x\,d\,vec\,\Sigma_x^{-1} \tag{B.26}$$

[47] Remember that the operator *vec* when applied to scalar or column vector, leads to the scalar or column vector itself.

Now, as,

$$d\Delta x = -d\,\underline{\mu}_X \tag{B.27}$$

and from (B.24),

$$\begin{aligned} d\left(\Delta x^T \Sigma_x^{-1} \Delta x\right) = & -\Delta x^T (\Sigma_x^{-1} + \left(\Sigma_x^{-1}\right)^T) d\,\underline{\mu}_X + \\ & -2\Delta x^T \otimes \Delta x \left[\left(\Sigma_x^T\right)^{-1} \otimes \Sigma_x^{-1}\right] N_n (A^* \otimes I_n)\left(\Delta^{1/2*}\right)^T \otimes I_n \left(A_0^T \otimes I_n\right) \tilde{\mathbf{G}}\, d\underline{\alpha} - \\ & -2\Delta x^T \otimes \Delta x \left[\left(\Sigma_x^T\right)^{-1} \otimes \Sigma_x^{-1}\right] N_n (A^* \otimes I_n)\left(I_p \otimes A\right) \mathbf{T} d\underline{\lambda} \end{aligned} \tag{B.28}$$

Therefore,

$$\begin{aligned} & D_{\underline{\mu}_X}\left(\Delta x^T \Sigma_x^{-1} \Delta x\right) = -\Delta x^T \left(\Sigma_x^{-1} + \left(\Sigma_x^{-1}\right)^T\right) \\ & D_{\underline{\alpha}}\left(\Delta x^T \Sigma_x^{-1} \Delta x\right) = -2\Delta x^T \otimes \Delta x \left[\left(\Sigma_x^T\right)^{-1} \otimes \Sigma_x^{-1}\right] N_n (A^* \otimes I_n)\left(\Delta^{1/2*}\right)^T \otimes I_n \left(A_0^T \otimes I_n\right) \tilde{\mathbf{G}} \\ & D_{\underline{\lambda}}\left(\Delta x^T \Sigma_x^{-1} \Delta x\right) = -2\Delta x^T \otimes \Delta x \left[\left(\Sigma_x^T\right)^{-1} \otimes \Sigma_x^{-1}\right] N_n (A^* \otimes I_n)\left(I_p \otimes A\right) \mathbf{T} \end{aligned} \tag{B.29}$$

B.1.5 Derivation of gradients for Λ (3)

We are now ready to derive the expressions for the gradients of the log-likelihood function Λ. Using the results derived so far, and the linearity of the gradient operators:

$$
\begin{aligned}
D_{\underline{\mu}_X}\Lambda(\underline{\mu}_X,\underline{\lambda},\underline{\alpha}) &= \frac{1}{2}\sum_{k=1}^{n_{obs}}\left\{\Delta x(k)^T\left(\Sigma_x(k)^{-1}+\left(\Sigma_x(k)^{-1}\right)^T\right)\right\}\\
D_{\underline{\lambda}}\Lambda(\underline{\mu}_X,\underline{\lambda},\underline{\alpha}) &= -\frac{1}{2}\sum_{k=1}^{n_{obs}}\left\{2\left\{vec\left(\Sigma_x(k)^{-1}\right)^T\right\}^T N_n(A^*\otimes I_n)\left(I_p\otimes A\right)\mathbf{T}\right\}+\\
&\quad+\frac{1}{2}\sum_{k=1}^{n_{obs}}\left\{2\Delta x(k)^T\otimes\Delta x(k)\left[\left(\Sigma_x(k)^T\right)^{-1}\otimes\Sigma_x(k)^{-1}\right]N_n(A^*\otimes I_n)\left(I_p\otimes A\right)\mathbf{T}\right\}\\
D_{\underline{\alpha}}\Lambda(\underline{\mu}_X,\underline{\lambda},\underline{\alpha}) &= -\frac{1}{2}\sum_{k=1}^{n_{obs}}\left\{2\left\{vec\left(\Sigma_x(k)^{-1}\right)^T\right\}^T N_n(A^*\otimes I_n)\left[\left(\Delta^{1/2*}\right)^T\otimes I_n\right]\left(A_0{}^T\otimes I_n\right)\tilde{\mathbf{G}}\right\}+\\
&\quad+\frac{1}{2}\sum_{k=1}^{n_{obs}}\left\{2\Delta x(k)^T\otimes\Delta x(k)\left[\left(\Sigma_x(k)^T\right)^{-1}\otimes\Sigma_x(k)^{-1}\right]N_n(A^*\otimes I_n)\left[\left(\Delta^{1/2*}\right)^T\otimes I_n\right]\left(A_0{}^T\otimes I_n\right)\tilde{\mathbf{G}}\right\}
\end{aligned}
\tag{B.30}
$$

Appendix C. Alternative HLV-MSPC Monitoring Procedures

The HLV-MSPC statistics presented in Chapter 7 lead to a direct calculation of the upper control limits to be used (no lower control limits are necessary), which are, for a given significance level (α) : $\chi_\alpha^2(n)$ for T_w^2 and $\chi_\alpha^2(n-p)$ for Q_w, where $\chi_\alpha^2(\nu)$ represents the upper $\alpha\times100\%$ percentiles for the χ^2 distribution with ν degrees of freedom. These types of limits are quite convenient, because they remain constant along time, despite the possible erratic variation of measurement uncertainties, but rely on assumptions regarding the probability density functions describing the behaviour of the random variables. In this context, a non-parametric approach for estimating the probability density function underlying T_w^2 and Q_w can be adopted as an alternative.

There are various ways for performing nonparametric density estimation, like those falling under the class of Kernel approaches, which estimate the underlying distribution using expressions of the form:

$$\hat{f}(t) = \frac{1}{n_{obs}} \sum_{i=1}^{n_{obs}} K\left(\frac{t - x_i}{h}\right) \tag{C.1}$$

where $\hat{f}$ represents the estimate of the true density f and $K(\cdot)$ is the kernel function satisfying,

$$\int_{-\infty}^{+\infty} K(t)\,dt = 1 \tag{C.2}$$

and h is the window width, a design parameter that should be adjusted in order to find the best compromise between smoothing and fit. Another class of approaches, developed from a quite different perspective, comprises those based on *orthogonal series estimators*. Basically, these techniques estimate f through an orthogonal series expansion, properly truncated to achieve a desired smoothing degree:

$$\hat{f}(t) = \sum_{i=-\infty}^{+\infty} \hat{c}_i \theta_i(t) \cong \sum_{i=k_1}^{k_2} \hat{c}_i \theta_i(t) \tag{C.3}$$

where $\{\theta_i\}_{i=-\infty}^{+\infty}$ is an orthonormal basis of the space under consideration. Some mild *a priori* assumptions are made regarding the nature of f, such as:

$$\int_{-\infty}^{+\infty} f^2(t)\,dt = k,\ k \text{ finite} \tag{C.4}$$

The c_i coefficients are given by,

$$c_i = \int_{-\infty}^{+\infty} \theta_i(t) f(t)\,dt \tag{C.5}$$

and a natural (unbiased) estimator for c_i, is:

$$\hat{c}_i = \frac{1}{n_{obs}} \sum_{k=1}^{n_{obs}} \theta_i(x_k) \tag{C.6}$$

There are still other approaches to nonparametric density estimation, like those arising from further developments to the kernel density estimation methods, such as the *nearest neighbour method*, or the *maximum penalized likelihood estimators* (Silverman, 1986). We will only mention here the approach based in estimating the probability density through an orthogonal polynomial series expansion, orthogonalized with respect to a given standard distribution (Tørvi and Hertzberg, 1997), that we will designate as "Pugachev" (following the reference provided by these authors):

$$\hat{f}(t) = f_s(t)\sum_{i=0}^{\infty} c_i p_i(t) \tag{C.7}$$

where the coefficients can be expressed as functions of the moments of the distribution.

In order to estimate the underlying distribution for the T_w^2 and Q_w statistics several nonparametric methods were tried, including: histograms, gaussian kernel density estimators (GKDE) and also, because of its easy implementation, Pugachev's approach. To illustrate their performance, we present some results, obtained when the techniques were applied to 2048 data values generated using the model that we will describe for the first case study in the next section. Figure C-1 represents the results obtained for T_w^2. In this figure, it is possible to see that both the Gaussian kernel density estimator (GKDE) and the Pugachev's method do not seem very adequate, since they provide estimates with a considerable coverage in the region of negative values. On the other hand, the histogram does provide an acceptable fit.

Thus, we did develop a methodology belonging to the class of orthogonal series estimators, based on wavelet basis functions. This methodology consists of constructing a function expansion such as (C.3), using orthogonal wavelet basis functions, and selecting the wavelet coefficients, obtained from (C.5), as a way to truncate it. For that purpose, we will simply neglect the detail coefficients, retaining only the approximation coefficients. We found by visual inspection that this procedure provides estimates with an adequate smoothing degree. Applying this methodology to the same data used in the generation of Figure C-1, the results presented in Figure C-2 were obtained, where it is

possible to see that the previous problems regarding non-zero probability for negative values have been overcome.

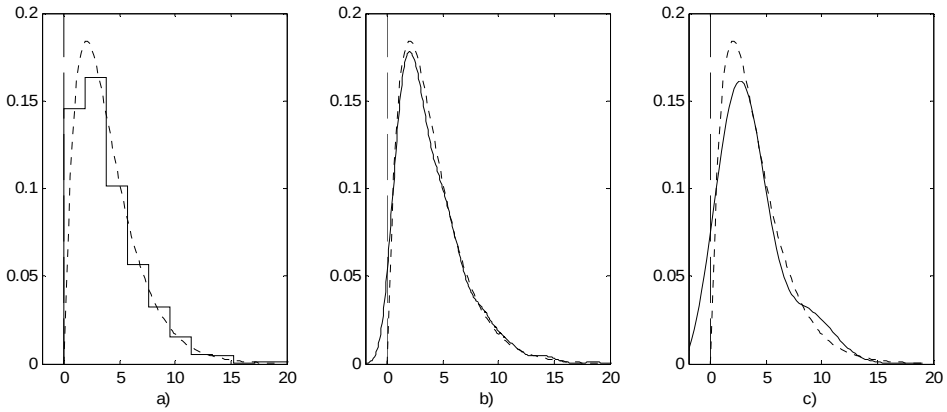


Figure C-1. Application of non-parametric density estimation techniques to simulated data: a) Histogram; b) Gaussian kernel density estimate; c) Pugachev's approach (solid lines). Dashed lines represent the expected χ^2 distribution of the T_w^2 statistic.

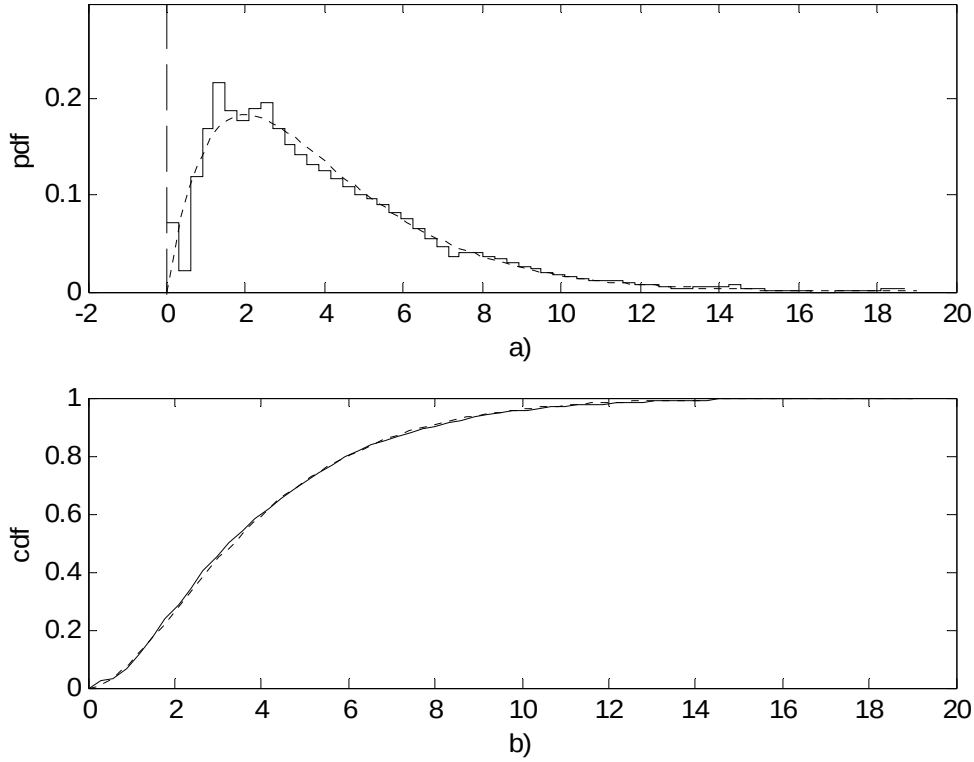


Figure C-2. Application of non-parametric wavelet density estimation techniques to simulated data: a) estimated probability density function, pdf (solid) and expected χ^2 distribution (dashed); b) cumulative distribution function, cdf (solid) and respective expected χ^2 distribution (dashed).

Therefore, we have now available a methodology that allows for a relaxation in the usage of parametric distributions for the T_w^2 and Q_w statistics, deriving alternative limits from historic data. There are still other alternative procedures for deriving control limits, that explore the availability of other types of information, e.g. repeatability and reproducibility (R&R) studies for the specification of measurement uncertainty, or noise characteristics of sensors. All this information can be used in conjunction with statistical and numerical methods, such as those based on re-sampling, noise addition, analytical or numerical linearization, in order to derive the NOC regions. The methodologies based on re-sampling and noise addition normally encompass a high number of evaluations leading to the calculation of the desired statistic. On the other hand, techniques based on analytic linearization are quite cumbersome, given the nature of the objective function and parameters involved (vectors and matrices). We therefore present here a simple approach based on noise addition, that does not require the repetitive calculation of an estimate of the model, and still provides a way for establishing control limits for the statistics, within a reasonable CPU time (Table C-1). It *assumes* that a HLV model structure is available at the time of implementation, and that information is available regarding the underlying random components in this model, i.e., about the latent variables and measurement noise random behaviour. As all projections of observed $x(k)$ onto the latent variable subspace will be always contaminated with (possibly heteroscedastic) noise, the calculation of the homoscedastic latent variable values, $l(k)$, is not possible, and therefore we can not rely on empirical probability distribution descriptions to characterize the isolated random behaviour attributed to latent variables. The same is not true for measurement noise, which can now have any probability distribution. Thus, in comparison with the non-parametric approach referred above for calculating control limits, this methodology not only allows for a relaxation in the usage of parametric distributions for T_w^2 and Q_w, but also enables the use of all available knowledge regarding the nature of measurement noise.

Table C-1. An alternative procedure for setting control limits in HLV-MSPC monitoring.

1. Identify the HLV and retain its structural part, and the distribution of the latent variables contained in the stochastic part;
2. For each new multivariate observation available, *k*:
    i. For i=1:n_{sim}: simulate the overall random process underlying $x(k)$ n_{sim} times, using the identified distribution for the random behaviour of the latent variables, and the distributions associated with measurement noise, through noise addition, or, if enough data available, using bootstrap, up-dated for time *k*:
        a. Calculate the T_w^2 and Q_w statistics for simulation *i*
    ii. Estimate the distributions of T_w^2 and Q_w at time *k* using data from the n_{sim} simulations;
    iii. Conduct a non-parametric one-sided hypothesis test to the observed $T_w^2(k)$ and $Q_w(k)$, using the distributions for these statistics at time *k*, and decide about their statistical significance, for a given α;
    iv. $k \to k+1$, Go To 2;
3. End

Figure C-3 illustrates the results obtained though the application of the above procedure to a model where all the underlying distributions of measurement noises are constant distributions (thus, not Gaussian). The limits are calculated both from a parametric approach, which assumes that all noise distributions are normal with zero mean and standard deviation calculated from the data, as well as from the alternative approach, where this assumption was relaxed. As can be seen, parametric statistical limits still do a good job in this case, which indicates a certain robustness to deviations from normality assumptions.

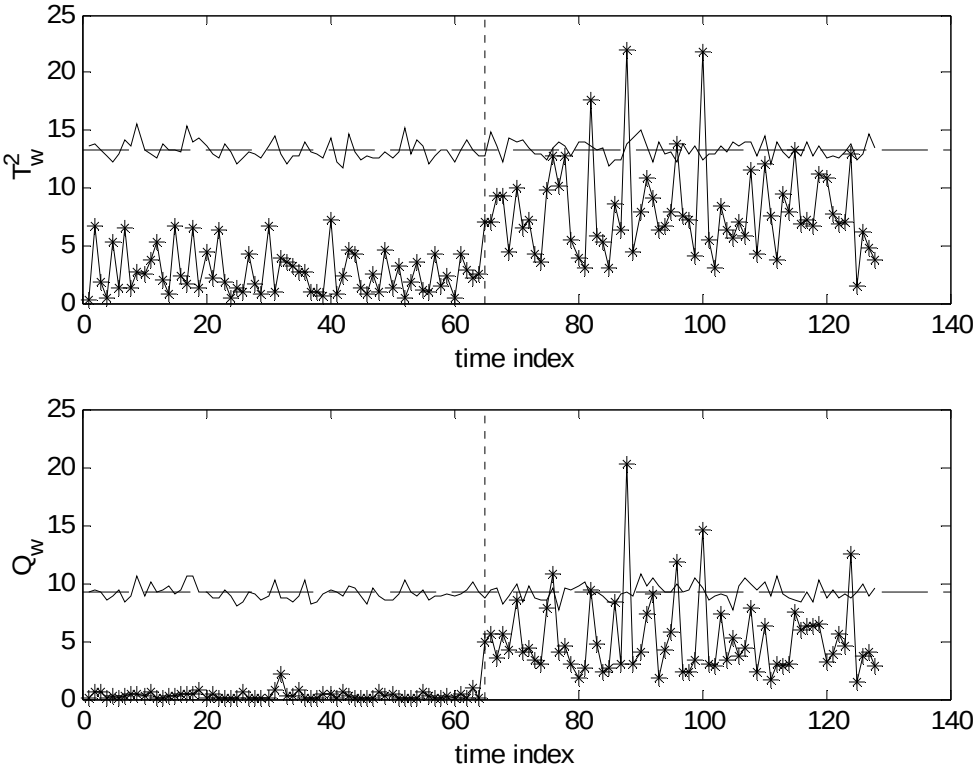


Figure C-3. HLV-MSPC results obtained with statistical limits calculated both from parametric assumptions (dashed line) and noise addition (solid line). The vertical dashed lines separate the test data in two regions: the first one regards normal operation and the second one reflects a step perturbation.

Appendix D. Principal Components Analysis (PCA)

PCA (Jackson, 1991; Johnson & Wichern, 1992; Martens & Naes, 1989) is a well known multivariate data analysis technique that addresses the problem of finding a reduced (p-dimensional) set of new variables, the principal components, which are linear combinations of the original (m) variables, with the ability of explaining most of their variability. Such linear combinations are those that successively present maximal (residual) variability (when the coefficients are constrained to unit norm), after the portion explained by the former components has been removed. The solution of such an optimization formulation can be reduced to an eigenvalue problem (Johnson & Wichern, 2002), where the optimal linear combinations (loadings) are given by the successive normalized eigenvectors of the data covariance matrix, associated with the eigenvalues sorted in a decreasing order of magnitude: the first principal component is given by the linear combination of the original variables provided by the eigenvector associated with the highest magnitude eigenvalue, etc.. Therefore, by applying PCA to the original data matrix, a set of correlated variables is transformed into a smaller, decorrelated one (i.e., having a diagonal covariance matrix), that often still explains a large part of the structure and variability present in the original data. The loadings are usually gathered in the columns of the $m \times p$ loading matrix, L, and the principal component values, or scores, appear in the $n \times p$ score matrix (n is the number of observations), T, leading to the following decomposition of the original data matrix:

$$X = TL^T + E \tag{D.1}$$

where E is a $n \times m$ residual matrix, which is in general a non-zero matrix when $p < m$, being the $\mathbf{0}$ matrix when $p = m$.

Regarding applications, PCA is recognized as being very effective on conducting several tasks (Jackson, 1991), such as dimensionality reduction, where the goal is to analyse data projected onto a lower dimensional subspace, without disregarding any variable or set of variables, being also very useful in developing visualization tools for detecting outliers, clusters, and in the interpretation of structural relationships among variables (Jolliffe, 2002). It can also be used in the context of regression analysis, where the uncorrelated linear combinations of input variables (principal components) become the new set of predictors, onto which the response is to be regressed (Jackson, 1991; Martens & Mevik, 2001; Martens & Naes, 1989), or in quality control (Jackson, 1959; Kresta *et al.*, 1991; MacGregor & Kourti, 1995), where the principal components become the relevant variables to monitor, along with the distance from each observation to the PCA subspace.

Appendix E. Mathematical Model for the Non-Isothermal CSTR under Feedback Control

The CSTR mathematical model supporting the simulations carried out in Section 9.4.4 is shown below (Luyben, 1990), according to the nomenclature, steady state and parameter values presented in Table E-1.

Global mass balance to CSTR

$$\frac{dV}{dt} = F_0 - F \tag{E.1}$$

Partial mass balance to component A

$$\frac{dVC_A}{dt} = F_0 C_{A0} - FC_A - k_0 e^{-E/RT} C_A V \tag{E.2}$$

Global CSTR energy balance

$$\frac{dVT}{dt} = F_0 T_0 - FT - \frac{\Delta H}{\rho C_p} k_0 e^{-E/RT} C_A V - \frac{UA}{\rho C_p}(T - T_{cj}) \tag{E.3}$$

Global cooling jacket energy balance

$$\frac{dV_{cj}T_{cj}}{dt} = F_{cj}(T_{cj,0} - T_{cj}) + \frac{UA}{\rho_j C_{p,cj}}(T - T_{cj}) \quad \text{(E.4)}$$

Control of reacting mixture volume (reactor level) using outlet flow rate

$$F = F_{set} - K_{c2}\left(V_{set} - V\right) \quad \text{(E.5)}$$

Control of CSTR temperature using cooling water flow rate

$$F_{cj} = F_{cj,set} - K_{c1}\left(T_{set} - T\right) \quad \text{(E.6)}$$

Table E-1. Variables used in the mathematical model and their steady state values, along with the model parameter values.

Variable / Parameter ↓ ↳	***Description***	***Steady state value / parameter value***
F	Outlet flow rate	40 ft^3h^{-1}
V	Reacting mixture volume	48 ft^3
C_{A0}	Concentration of reactant *A* in the inlet stream	0.5 lb·mol A ft^{-3}
C_A	Concentration of reactant *A* in the CSTR	0.245 lb·mol A ft^{-3}
T	Temperature in the CSTR	600 ºR
T_{cj}	Temperature in the cooling jacket	594.6 ºR
F_{cj}	Water flow in the cooling jacket	49.9 ft^3 h^{-1}
T_0	Temperature in the inlet stream	530 ºR
V_{cj}	Cooling jacket volume	3.85 ft^3
k_0	Pre-exponential factor	7.08×10^{10} h^{-1}
E	Activation energy	30 000 Btu $lb{\cdot}mol^{-1}$
R	Gas constant	1.99 Btu $lb{\cdot}mol^{-1}$ $ºR^{-1}$
U	Overall heat transfer coefficient	150 Btu h^{-1} ft^{-2} $ºR^{-1}$
A	Heat transfer area	250 ft^2
$T_{cj,0}$	Temperature in the cooling jacket's inlet stream	530 ºR
ΔH	Heat of reaction	-30 000 Btu $lb{\cdot}mol^{-1}$
Cp	Heat capacity of the mixture	0.75 Btu lb_m^{-1} $ºR^{-1}$
ρ	Density of the mixture	50 lb_m ft^{-3}
$C_{p,cj}$	Heat capacity of the cooling liquid (water)	1 Btu lb_m^{-1} $ºR^{-1}$
ρ_{cj}	Density of the cooling liquid (water)	62.3 lb_m ft^{-3}
K_{c1}	Tuning constant for the proportional action in the temperature control loop	4 ft^3h^{-1} $ºR^{-1}$
K_{c2}	Tuning constant for the proportional action in the level control loop	10 h^{-1}
F_{set}	Set point for outlet reactor flow	40 ft^3 h^{-1}
$F_{cj,set}$	Set point for cooling jacket flow	49.9 ft^3 h^{-1}
T_{set}	Set point for reactor temperature	600 ºR
V_{set}	Set point for reacting mixture volume	48 ft^3

Figure E-1 and Figure E-2 present the values for the 10 variables involved in the simulation of the CSTR dynamic behaviour, regarding the reference set used in Section 9.4.4.

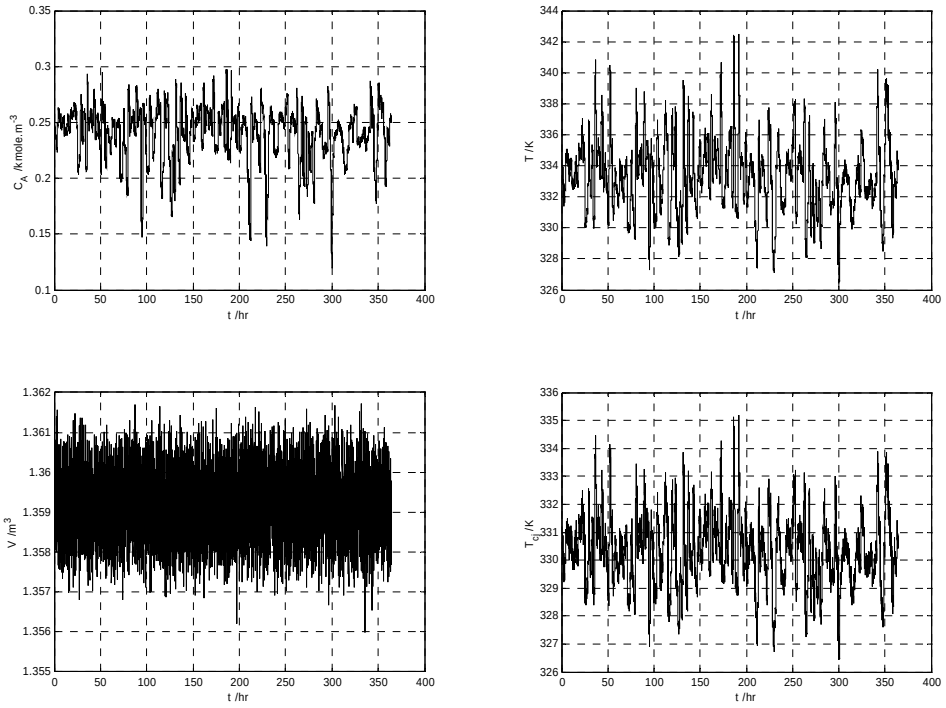


Figure E-1. Values for $\{C_A, T, V, T_{cj}\}$ in the reference data set.

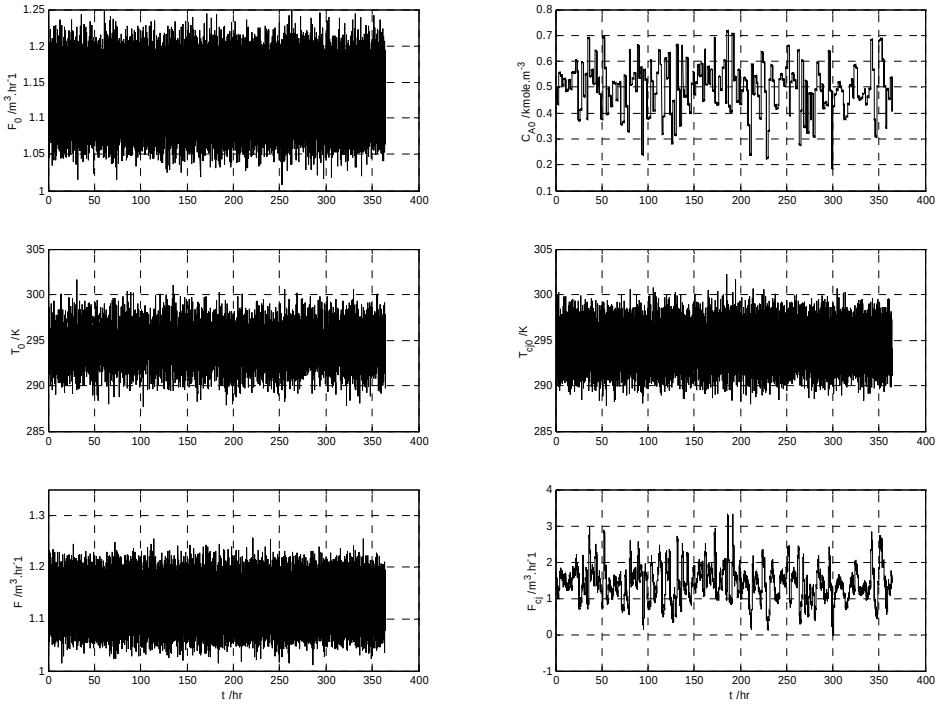


Figure E-2. Values for $\{F_0,\ C_{A0},\ T_0,\ T_{cj,0},\ F,\ F_{cj}\}$ in the reference data set.