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

DESENVOLVIMENTO DE MODELOS QUIMIOMÉTRICOS UTILIZANDO O ALGORITMO DE OTIMIZAÇÃO COLÔNIA DE FORMIGAS

TESE DE DOUTORADO

Cassiano Ranzan

Porto Alegre

2014

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

DESENVOLVIMENTO DE MODELOS QUIMIOMÉTRICOS UTILIZANDO O ALGORITMO DE OTIMIZAÇÃO COLÔNIA DE FORMIGAS

CASSIANO RANZAN

Tese de Doutorado apresentada como requisito parcial para obtenção do título de Doutor em Engenharia

Área de concentração: Pesquisa e Desenvolvimento de Processos

Linha de Pesquisa: Engenharia de Sistemas – Projeto, Modelagem, Controle e Otimização de Processos

Orientador:
Prof. Dr. Jorge Otávio Trierweiler

Co - Orientador:
Prof. Dr. Luciane Ferreira Trierweiler

Co – Orientador Estrangeiro:
Prof. Dr. Bernd Hitzmann

Porto Alegre

2014

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

A Comissão Examinadora, abaixo assinada, aprova a Tese "Desenvolvimento de Modelos Quimiométricos Utilizando o Algoritmo de Otimização Colônia de Formigas", elaborada por Cassiano Ranzan, como requisito parcial para obtenção do Grau de Doutor em Engenharia.

Comissão Examinadora:

Prof. Dr. Michel J. Anzanello – UFRGS

Prof. Dr. Antonio J. G. Cruz – UFSCAR

Prof. Dr. Marco Ferrão – UFRGS

Resumo

O objetivo principal da otimização de processos é a obtenção de alta produtividade e lucro, seja em processos químicos ou bioquímicos. Para alcançar isto, se faz necessária a utilização de técnicas e estratégias de controle, individualizadas e apropriadas para cada situação. Uma vez que se dispõe de atuadores e manipuladores de variáveis para as mais diversas aplicações (desde extremos de temperatura e pressão, até casos onde as características do meio o tornam inapropriado à utilização de sondas), a restrição na utilização de controle avançado recai sobre o déficit de coleta e caracterização dos estados do processo frente à suas variáveis de interesse. Muitas vezes, as características do meio inviabilizam a utilização de sensores, como em casos com alto risco de contaminação. Também existem os casos onde apenas técnicas laboratoriais são viáveis para a caracterização de certas variáveis, e nestes casos, usualmente, o tempo de resposta dos métodos é muito elevado, limitando a ação de controle e impossibilitando a operação do processo em ramos de maior produtividade. Neste contexto, sensores óticos associados a métodos quimiométricos são considerados a evolução natural das técnicas para caracterização de processos. Baixo tempo morto, medidas não intrusivas, baixo custo, simples manutenção, alta sensibilidade e seletividade, são algumas das características inerentes a este tipo de sensor. Neste trabalho, são estudadas e associadas técnicas de espectroscopia vibracional (Infravermelho Próximo e Fluorescência 2D) com metodologias quimiométricas baseadas em elementos espectrais puros. É apresentada a metodologia de caracterização denominada Modelagem Quimiométrica com Componentes Espectrais Puros (PSCM), onde modelos multilineares são ajustados para predição de determinadas variáveis de interesse, através de elementos constituintes de medidas de espectroscopia, selecionados com a utilização do método de otimização de Colônia de Formigas. Como casos de estudo são utilizados o procedimento de determinação do teor de proteínas em amostras de farinha de trigo e centeio utilizando Espectroscopia de Infravermelho Próximo (NIR), e a caracterização das variáveis de estado do processo fermentativo de glicose com a levedura *Saccharomyces cerevisiae*, através de medidas de Espectroscopia de Fluorescência 2D. A metodologia apresentada é comparada com as técnicas quimiométricas usuais (Análise de Componentes Principais (PCA), Regressão de componentes Principais (PCR), Regressão por Mínimos Quadráticos Parciais (PLSR), Regressão PLS com Janela Móvel (MWPLSR) e PLS com Janela Móvel Variada (CSMWPLS)), permitindo incrementos significativos na qualidade de predição das variáveis de interesse, para ambos os processos, além de apresentar uma nova opção para a seleção e filtragem de dados espectrais, baseada em informações qualitativas do ajuste de modelos, denominada trilha de feromônios.

Abstract

The aim of process optimization is obtaining higher productivity and profit in chemical or bio-chemical process. For that, one must apply control techniques specially designed for each situation. Once there actuators are available for several applications (from extremes of temperature and pressure, even cases where the medium characteristics are not suitable for probes), the restriction on advanced control is due to the lack of measurements. In some applications, measurements are difficult because of high contamination risks. In other cases, only laboratory methods are available, resulting in big time delay, limiting automatic control and, therefore, operation on regions of high production and profit. At this context, optical sensors associated with chemometric methods are the natural evolution for process characterization tools. Low time delay, non-intrusive measurements, low costs, easy maintenance, high sensibility and selectivity, are some examples of the characteristics usually associated with those sensors. In this work, it is studied and associated techniques of vibrational spectroscopy (Near Infrared and 2D Fluorescence) with chemometric methodologies based on pure spectral elements. This work presents the method called Pure Spectral Chemometric Modeling (PSCM), where multi linear models are adjusted for infer specific interest variables through spectral elements, selected using the optimization algorithm ACO (Ant Colony Optimization). Study cases are: (i) determination of protein content amount presented by wheat and rye flour samples using Near Infrared; and, (ii) characterization of significant variables of Saccharomyces cerevisiae fermentative process with glucose as substrate, measured with 2D Fluorescence Spectroscopy. Presented methodology is compared with usual chemometric methodologies (Principal Component Analysis (PCA), Principal Component Regression (PCR), Partial Least Squares (PLS), Moving Window Partial Least squares Regression (MWPLSR) and Changeable Size Moving Window Partial Least Squares (CSMWPLS)), showing significant improvement on prediction, for both cases. The proposed method is also a new option for pre-selection of spectral data, only based on qualitative information of assisted modeling process, called pheromone trail.

"We live in a Multivariate World"

Juan Antonio Fernandez Pierna

Agradecimentos

Primeiramente gostaria de agradecer aos meus pais, Angelin Ranzan e Iracema Inêz Ceron Ranzan, por toda a ajuda e apoio, sem os quais eu jamais concretizaria esta etapa... eles são os grandes responsáveis por eu ser quem sou hoje.

Ao meu irmão, companheiro, colega de apartamento, de departamento, de grupo de pesquisa, de treino e daquilo mais que aparecer pelo caminho, Lucas Ranzan... valeu pela força em tudo mano!! A próxima vai ser a tua!!

À minha namorada Alina, uma mulher incrível, de princípios e objetivos claros, que soube entender minhas angustias e necessidades. Muito obrigado pelo constante apoio e incentivo, principalmente nas etapas finais deste trabalho. Gostaria de agradecer também à família Savaris, que me recebeu de braços abertos e proporcionou inúmeros momentos de alegria, inclusive, sendo responsável por me introduzir a um novo mundo (muitas vezes marginalizado pela cultura popular, mas que transcende o tempo e o conceito de folclore) o Tradicionalismo Gaúcho.

Aos Professores Dr. Jorge Otávio Trierweiler e Dr Luciane Ferreira Trierweiler, pela orientação, amizade e conselhos. Eles foram os idealizadores deste trabalho, meus mentores e grandes visionários da evolução das técnicas de caracterização de processos!

Ao Professor Dr. Bernd Hitzmann, que durante o período em que eu estive em seu departamento, me tratou com extrema atenção e prestatividade, sempre com dedicação e solicitude. Através da sua pessoa, gostaria de estender os agradecimentos a todo o grupo de colegas da Universität Hohenheim: Tetyana (e Amadeus, “Die Hund”), Marius, Bianca, Marc, Florian, Olivier, Annika, Viktoria, Lena, Melina e Saskia.

Ao Grupo de Intensificação, Modelagem, Simulação, Controle e Otimização de Processos (GIMSCOP), por me receber de braços abertos. Não poderia esquecer os colegas que me proporcionaram momentos únicos no decorrer destes anos de convivência, como o MSc. Eng. Químico Rodrigo Paulo Silveira, atualmente mais conhecido como “Agente Silveira”, mais novo integrante do corpo de agentes da PRF. Também aos colegas: Giovana, Jordana, Júlia, Fernanda, Juliana, João, Raimundo, Ana, Carol, Cris, Nelzy, Larissa, Rúbia, dentre muitos outros, pelos momentos de descontração e discussões.

Aos meus irmãos da Família Sonkão Camp, especialmente ao meu amigo Helio Dipp, que mais que um professor, é um exemplo de dedicação e amor em relação a tudo que faz e um exemplo de fibra moral. Ao Leoton “The Spartan” Costa, o campeão com as sobrancelhas mais bem cuidadas que já se viu por essas bandas. Ao Roberto, Leonardo, Renan, Igor, Matheus, Menegas, Tiagão, Ivan, Peter, Luciano e demais, pela amizade, parceria e acima de tudo, pelo incontestável apoio em todas as horas.

Aos meus eternos amigos de Caxias, Chorão, Ede, Paulo, Muller, Fontana, Franja, Léo, Cléberson e Polleto, pela força, sempre demonstrando interesse em saber quando eu vou começar a “trabalhar de verdade” e parar de “só estudar”.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

ABREVIAÇÕES

ACO	Otimização Colônia de Formigas (*Ant Colony Optimization*)
CSMWPLS	Mínimos Quadráticos Parciais de Janelas Móveis Variáveis (*Changeable Size Moving Window Partial Least Squares*)
LV	Vetor de Carga (*Load Vector*)
MISO	Entradas Múltiplas e Saída Única (*Multiple Imput Single Output*)
modCSMWPLS	Mínimos Quadráticos Parciais de Janelas Móveis Variáveis Modificado (*Changeable Size Moving Window Partial Least Squares*)
MWPLSR	Regressão de Mínimos Quadráticos em Janelas Móveis (*Moving Window Partial Least Squares*)
NIR	Infravermelho Próximo (*Near Infrared*)
PC	Componente Principal (*Principal Component*)
PCA	Análise de Componentes Principais (*Principal Component Analysis*)
PCR	Regressão de Componentes Principais (*Principal Component Regression*)
PLS	Mínimos Quadráticos Parciais (*Partial Least Squares*)
PLSR	Regressão com Mínimos Quadráticos Parciais (*Partial Least Squares Regression*)
PSCM	Modelagem com Componentes Espectrais Puros *(Pure Spectral Component Modeling*)
R^2	Índice de Correlação
RMSE	Raiz Quadrada do Erro Quadrático Médio (*Root Mean Square Error*)
RMSEC	Raiz Quadrada do Erro Quadrático Médio da Calibração (*Root Mean Square Error of Calibration*)
RMSEP	Raiz Quadrada do Erro Quadrático Médio da Predição (*Root Mean Square Error of Prediction*)
SNV	Padrão Normal de Variação (*Standard Normal Variate*)
SSR	Soma dos Resíduos quadráticos (*Summed Square Residues*)

Capítulo 1 – Introdução

1.1 Motivação

Com o incremento da demanda por produtos industrializados, sem mencionar a competição econômica dentro dos diversos segmentos, exige-se que as empresas operem cada vez mais próximas ao nível máximo de desempenho permitido, seja para garantir vantagens financeiras e logísticas sobre seus concorrentes, seja para um melhor aproveitamento dos recursos disponíveis.

Com o objetivo de incrementar a produtividade dos processos, surge a necessidade de utilização de técnicas de otimização; técnicas estas que possibilitam o redimensionamento de processos e operações em direção ao seu nível máximo de eficiência. Assim, dentre os diversos benefícios associados à utilização de técnicas de otimização de processos, pode-se salientar: melhor alocação de recursos de produção e logística, utilização de potenciais sinergias entre elementos da cadeia produtiva, redução de custos de produção e aumento na eficiência do processo produtivo.

Estima-se que a indústria química brasileira foi responsável pela movimentação de cerca de R$ 293 milhões no ano de 2012. Esse valor é cerca de 2,7 vezes maior que a quantia movimentada por essa mesma indústria há apenas 17 anos, sendo atualmente considerada a sexta maior do mundo no ramo (de acordo com relatórios da ABIQUIM - Associação Brasileira de Indústria Química (2013)). A evolução do cenário da indústria química nas últimas décadas mostra a importância desse setor para a economia nacional, de forma que o aprimoramento dos diversos processos de produção deve acompanhar naturalmente esse crescimento.

Dentro da matriz de produção de insumos químicos, os processos biotecnológicos vêm assumindo cada vez mais espaço em relação a processos industriais usuais, não apenas devido a seu apelo ambiental, mas muito devido à disponibilidade de fontes alternativas de matéria-prima, muitas vezes apresentando facilidades operacionais ou de logística.

Historicamente, o caminho mais eficaz para a obtenção de incremento produtivo estava associado à evolução dos processos (Aynsley M, Hofland A *et al.* 1993). Entretanto, os últimos anos revelam significantes avanços em eficiência devido à aplicação de conceitos da área de supervisão e controle, já que estas técnicas permitem reduções nos custos de produção e incremento na conversão das reações, além de garantirem a qualidade do produto desejado (Yamuna and Ramachandra 1999).

Analogamente aos processos químicos, o incremento produtivo de bioprocessos está associado ao uso de ferramentas de controle. Do ponto de vista operacional, o controle de processos implica na necessidade de caracterização dos estados do sistema. O grande problema associado à bioprocessos e processos químicos se encontra justamente nessa etapa, que necessita, dentre outras características, ser rápida e precisa (Lindemann, Marose *et al.* 1998, Ranzan, Trierweiler *et al.* 2011).

Atualmente, quando se trata de sistemas de controle computacional, existem comercialmente diversas plataformas para análise de dados e produtos, geralmente representativos para variações genéricas da estrutura geral dos processos (Alford 2006).

Em processos químicos ou bioquímicos certas propriedades são de difícil medição, seja pelos custos proibitivos de analisadores em linha ou pela inexistência de técnicas de determinação *on-line* (Diehl, Secchi *et al.* 2009). A utilização de medições laboratoriais, ou cromatográficas, acarreta um elevado atraso de tempo para a realimentação de informações, podendo limitar consideravelmente o desempenho alcançável pelo sistema de controle. Isso porque os controladores requerem uma realimentação mínima de informações do processo. Para minimizar esse problema de realimentação deve ser considerado principalmente o desenvolvimento de novos sensores, a melhora dos sistemas de amostragem e a utilização de técnicas de estimação *on-line* (Bequette 2003, Ranzan 2010)

A fim de se obter maior produtividade e qualidade de produtos constante, bem como permitir a utilização de ferramentas de otimização e controle apropriadas, é necessário o monitoramento em tempo real dos parâmetros chave das reações, como concentrações de reagentes, produtos, biomassa, substrato e/ou dados de viscosidade, pH, temperatura, dentre outras propriedades físico-químicas do meio reacional (Surribas, Geissler *et al.* 2006).

Controle de processos é definido como sendo o provedor do meio reacional ideal para os reagentes (ou microrganismos no caso de bioprocessos) de forma a favorecer a obtenção de um determinado produto desejado. Esta função inclui disponibilizar quantidade certa de reagentes ou nutrientes para a cultura (p.ex. carbono, nitrogênio, oxigênio, fósforo, enxofre, dentre outros minerais), remoção de qualquer subproduto indesejado e manipulação de parâmetros reacionais importantes (por exemplo, temperatura e pH).

Além do ponto de vista prático, a demanda por técnicas de monitoramento rápido de processo favorecem a obtenção de melhores modelos de processo, melhor controle das variáveis controladas, desenvolvimento de operadores independentes, robustos, padronizados e transferíveis; tudo que aumenta significativamente o entendimento acerca dos sistemas reais e permite o estudo e desenvolvimento de metodologias de operação individualizadas (Whitford and Julien 2007).

A capacidade em desenvolver estratégias de controle e otimização de processos está diretamente relacionada com a disponibilidade de sensores *in situ*, ou à instrumentação analítica em tempo real (*Real-Time Analytical Instrumentation*) disponível para cada sistema, sendo esse assunto identificado pelo Conselho Nacional de Ciência e Tecnologia em 1995 como prioritário na destinação de fundos de pesquisa de uma forma genérica (Science. 1995). A demanda por técnicas de monitoramento rápido de processos é voltada para:

- Obtenção de maiores informações sobre processos, permitindo o desenvolvimento de modelos mais representativos;
- Sistemas de controle mais eficientes e precisos na manipulação da taxa alimentação e demais parâmetros significativos do processo;
- Necessidade de obtenção de sistemas robustos, padronizados, transferíveis e de operação independente;

Técnicas de monitoramento, assim como os sensores e metodologias analíticas a elas associadas, podem ser classificadas de acordo com o nível de contato com o meio, variando desde medidas *off-line* e/ou *at-line* (fazem uso de instrumentação analítica) até medidas *on-line*. Diversas configurações *on-line* estão atualmente disponíveis, desde a inserção direta de sensores *in-situ* até *ex-situ* com loops de amostragem (Whitford and Julien 2007). A Figura 1.1 apresenta a classificação de sensores com relação ao respectivo nível de invasividade apresentado por cada sensor ao processo.

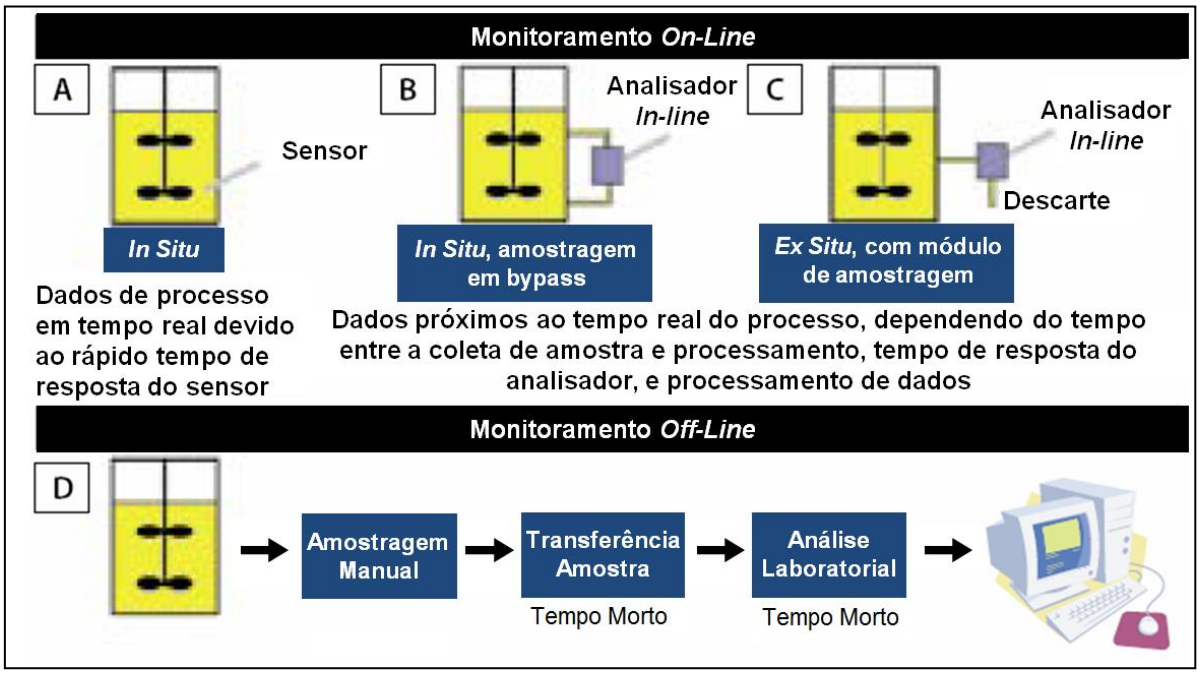

Figura 1.1: Classificação de monitoramento de processos, sensores e configurações de análises em função do grau de invasão. **Fonte: Adaptado de Whitford e Julien (2007).**

O estado da arte em monitoramento de processos requer que cada parâmetro monitorado em um sistema de controle possa ser obtido através do emprego de um pequeno sensor eletrônico *in situ*, sendo que esse sensor deve ser capaz de satisfazer certos requisitos como:

- Ser inerte ao processo;
- Ser passível de esterilização no local ou autoclavagem;
- Permitir limpeza no local, além de ser facilmente lavados manualmente;
- Possuir sensibilidade e resolução suficientes, necessárias ao sistema de monitoramento e controle;
- Ser facilmente calibrável através da utilização de soluções e procedimentos padrões aplicados em laboratórios;
- Apresentar pouco ruído e baixa interferência causada por fatores externos;
- Ser robusto para operar nas condições de processos;
- Apresentar alta relação benefício/custo.

Com base nas características almejadas pelo sensor ideal, vem surgindo um crescente interesse envolvendo o desenvolvimento e aplicação de sensores ópticos na caracterização de processos. A rápida evolução desse tipo de sensor, aparentemente muito promissor para a aplicação nos mais variados ramos de processos químicos e bioquímicos, aliada às suas características e funcionalidades, tem chamado a atenção dos pesquisadores. A atratividade de sensores ópticos em relação a sensores convencionais reside nas vantagens de sua utilização, principalmente sua imunidade eletromagnética às interferências, permitindo a construção de equipamentos não invasivos, de monitoramento remoto, contínuos, com tamanho reduzido e capazes de mensurar simultaneamente diversas grandezas de interesse (Skibsted, Lindemann *et al.* 2001, Wong, Chan *et al.* 2014).

Inúmeras pesquisas vêm sendo realizadas acerca da caracterização de processos utilizando técnicas baseadas em princípios ópticos, dentre as quais se pode citar os trabalhos de Rhee *et al* (2004) onde os autores estudam o monitoramente e controle *on-line* da concentração de substrato em processos biológicos por sistemas de injeção, Schügerl (2001) que apresenta um apanhando do progresso das técnicas de monitoramento, modelagem e controle de bioprocessos nos últimos 30 anos, Wong *et al* (2014) que estudou o desenvolvimento de um sensor em miniatura de fibra óptica para a determinação do pH em reações, o trabalho de Mulchandani e Bassi (1995) que apresenta os princípios para a criação de sensores ópticos voltados ao monitoramento e controle de bioprocessos, o trabalho de Fernandes *et al*. (2008) que estudou a determinação simultânea de metanol e etanol na gasolina através de espectroscopia de NIR, e o trabalho de Killner, Rohwedder e Pasquini (2011) que aplicaram modelos de regressão PLS em dados de espectroscopia NIR para o monitoramento on-line da reação de produção de biodiesel. Quando tratamos de sensores ópticos, técnicas distintas estão associadas a essa classificação, incluindo espectroscopia de fluorescência, espectroscopia de infravermelho (NIR), ressonância plasmática de superfície, espectroscopia RAMAN e espectroscopia de interferência refletiva.

A Figura 1.2 apresenta um exemplo de aplicação industrial de sensores de processo *in-situ* baseados na técnica de NIR, sendo utilizados para acompanhamento do processo de produção de fibra de papel. Os sensores apresentados neste exemplo são da empresa Foss, interfaceados com o processo através de cabos de fibra ótica e sondas de refletância difusas, com varredura de comprimento de onda no intervalo 400 – 2200 nm.

No processo em questão, as condições operacionais são de 180 °C, 5 bar e 20 $m.s^{-1}$ (velocidade de escoamento da polpa) (Kessler and Kessler 2014).

Figura 1.2: Sensor NIR *In-line* de fibra ótica aplicado para o acompanhamento do processo de produção de fibra de papel. a) Etapa do processo fabril, b) sensor de espectroscopia NIR e c) sensor sendo aplicado diretamente no processo. **Fonte: Adaptado de Kessler e Kessler (2014).**

Espectroscopia de Infravermelho Próximo (NIR) é uma técnica utilizada há bastante tempo para o monitoramento de biomassa, glicerol, glicose, amônia, acetato e lactato em processos microbianos, sendo atualmente adaptada para o acompanhamento de culturas de células animais (McLeod, Clelland *et al.* 2009), determinação do número de octanos em gasolinas, concentração de glicose no sangue, concentração de óleos em sementes e graduação Brix em cana de açúcar (Riley, Okeson *et al.* 1999, Vaidyanathan, Arnold *et al.* 2001, Allegrini and Olivieri 2011), dentre outros.

Da mesma forma, sensores baseados em tempo de vida de fluorescência associam a intensidade de fluorescência apresentada por soluções para determinar a concentração de componentes do meio reacional, tendo grande apelo associado à geração de sensores em miniatura para biorreatores no formato *single-use* (Ge, Kostov *et al.* 2005). Novos sensores baseados nessa técnica já mostraram ser eficientes na quantificação de CO_2, com tempo de resposta de aproximadamente 6 minutos, além de já terem sido propostos sensores *on-line* para o monitoramento de NADPH, juntamente com outros fluoróforos biogênicos, como proteínas, coenzimas e vitaminas (Marose, Lindemann *et al.* 1998, Pattison, Swamy *et al.* 2000).

É importante salientar que não é toda informação coletada por sensores baseados em espectros que apresentam relevância para a caracterização do processo (Boehl, Solle *et al.* 2003), de forma que é essencial a utilização de analisadores virtuais ou estimadores de estado para traduzir o que é lido *on-line* pelo espectrômetro em informação útil à operação do sistema. Medidas de espectroscopia podem ser geradas em intervalos significativamente menores que a unidade de tempo padrão da maioria dos processos, dependendo basicamente da resolução dos espectros coletados.

O número de informações geradas por tais equipamentos é extremamente elevado. Se uma varredura completa de espectro é realizada a cada 10 minutos, são gerados aproximadamente 150.000 pontos de medida (o equivalente a 1,5 MB) em 24 horas. Para traduzir o elevado número de informações obtidas são empregados métodos estatísticos, isto é, modelos capazes de converter dados de espectroscopia em variáveis de interesse, tais como: composições, indicadores de qualidade, etc. (Solle *et al.*, 2003; Hitzmann *et al.*, 1998).

Métodos estatísticos clássicos são usualmente baseados na escolha de premissas assumidas para os modelos, como por exemplo, distribuições normais dos dados, ou então, independência das observações. Na prática, entretanto, essas premissas podem ser violadas ou mesmo invalidadas para o conjunto de dados analisados. Dados reais são compostos por pontos fora da tendência (*outliers*) que podem mascarar distribuições ordenadas, ou então, as distribuições podem não ser relacionadas de acordo com o modelo assumido. Nesse ponto, fica a dúvida: a aplicação de métodos estatísticos clássicos, baseados em premissas assumidas genericamente, é capaz de conduzir à conversão de dados de espectroscopia em informações de processo? Especialmente para grandes conjuntos de dados – como no caso de dados de espectroscopia que apresentam alta sensibilidade e correlação entre elementos espectrais, associada à elevada quantidade de informações – esta linha de ação não aparenta ser promissora, de forma que surge a necessidade de aplicação de técnicas de modelagem mais robustas, capazes de converter grandes matrizes de dados em variáveis condensadas que carregam consigo a maior quantidade possível de informação qualitativa e quantitativa contidas nas matrizes originais (Filzmoser and Todorov 2011).

Um das possibilidades para resolução desse problema é através do uso de técnicas de Modelagem Quimiométrica (*Chemometric Modeling*). Modelagem Quimiométrica pode ser definida como a aplicação matemática, estatística, simbólica ou gráfica de métodos para maximização de informações químicas as quais devem ser extraídas de um determinado conjunto de dados (ex. dados de espectroscopia).

Esse tipo de modelagem é usualmente aplicada em problemas de descrição e predição em dados de processos, particularmente com grande aplicabilidade em química analítica. Para aplicação em problemas descritivos, propriedades de sistemas químicos são modeladas com a intenção de obter-se o entendimento de relações fundamentais e estruturais dos sistemas (p.ex. identificação de modelos). De outra forma, características de sistemas químicos são modeladas com a intenção de predizer propriedades ou comportamentos de interesse. Em ambos os casos, esses modelos são voltados para aplicação em grandes conjuntos de dados, usualmente envolvendo milhares de variáveis com centenas de amostragens cada (Brereton 2013).

Quanto às técnicas de modelagem quimiométrica usualmente empregadas no tratamento de dados, podem ser classificadas em quatro grupos, de acordo com o objetivo de sua aplicação: Análise Exploratória, Regressão Multivariada, Classificação e Planejamento de Experimentos.

Do ponto de vista operacional, o objetivo central da caracterização e acompanhamento de processos reside na predição das variáveis de estado de interesse, de forma que as técnicas de quimiometria classificadas como de Regressão Multilinear são as que apresentam maior interesse desse segmento. Nesse sentido, Modelos Multilineares são construídos para correlacionar espectros de amostras de referência com suas, previamente conhecidas, variáveis de estado. Esses modelos são construídos de forma a fornecer os valores das propriedades de referência como função da matriz de espectros coletada. A relação entre as propriedades e os espectros é expressa através da chamada matriz de coeficientes de regressão, que pode ser obtida através dos métodos previamente citados. Desta forma, modelos multivariados são constituídos por um dado número, fixo, de variáveis de entrada e parâmetros de modelo, sendo a seleção dessas características uma função considerada não trivial. O foco da seleção de variáveis é a obtenção de modelos baseados em dados espectrais que carreguem consigo o máximo de informação relativa aos compostos ou propriedades de interesse (Gomes, Galvão *et al.* 2013).

Existem, de forma genérica, dois tipos de métodos de seleção de variáveis: (1) Inspeção espectral completa dos coeficientes de regressão (Regressão por mínimos Quadráticos Parciais, Regressão por componentes principais, etc.) ou dos vetores de expressão (vetores de carga, componentes principais, etc.), e (2) Busca por faixas de espectro nas quais o erro de predição é mínimo. O método mais simples e ainda muito utilizado por pesquisadores é a inspeção visual de espectros, onde variáveis cujo vetor de regressão é significante são incluídas nos modelos, enquanto aquelas ruidosas ou cujo vetor de regressão apresenta baixa intensidade, não (Andries, Heyden *et al.* 2013).

A busca por regiões do espectro onde os indicadores para predição podem ser considerados os melhores possíveis constituem uma alternativa válida para seleção de variáveis. A seleção de regiões espectrais específicas é considerada a escolha de janelas espectrais que possuem a maior quantidade de informação a respeito do analito de interesse. Um dos métodos que atua neste sentido é chamado Regressão por Mínimos Quadráticos Parciais por Intervalo (i-PLS) (Jiang, Berry *et al.* 2002). Esse método constrói um modelo multivariado em cada uma das janelas do espectro, selecionadas através da estratégia de movimento de uma janela de tamanho fixo. Desta forma, a melhor região espectral é considerada aquela com o menor erro de predição dentre as janelas selecionadas (Filgueiras, Alves *et al.* 2014). Uma evolução desta técnica emprega janelas com tamanhos variados, com o indicador de erro sendo dependente da predição da janela e do seu tamanho da mesma. Essa variação permite a seleção de janelas com tamanho superior ao tamanho mínimo escolhido, entretanto, não pode localizar regiões através da combinação de sub-regiões separadas (Olivieri, Goicoechea *et al.* 2004).

Já que a busca exaustiva pode ser proibitiva devido ao tempo computacional requerido, principalmente para espectros com alta resolução e faixa de medida elevada (o que implica em grande número de componentes espectrais em cada amostragem), estratégias alternativas estão sendo propostas para a seleção de grupos de elementos espectrais, baseadas em algoritmos para busca de ótimos globais do tipo heurísticos

(inspirados em processos naturais, p.ex. Algoritmos Genéticos) (Leardi, Seasholtz *et al.* 2002).

Recentemente, o método de Otimização de Colônia de Formigas (*Ant Colony Optimization* - ACO) tem sido introduzido para seleção de variáveis em problemas de regressão PLS. O algoritmo ACO é inspirado no comportamento da colônia de formigas em busca do melhor caminho entre as possíveis fontes de comida e seu formigueiro. As variáveis de escolha são associadas aos possíveis caminhos espaciais utilizados pelas formigas, de forma que são "marcados" de acordo com sua eficiência na predição da grandeza almejada. Um dado caminho é conectado a um conjunto de variáveis selecionadas, que correspondem a um erro de predição. Em cada iteração, as formigas depositam uma dada quantidade de Feromônio, que aumenta com a diminuição do erro da função objetivo. As formigas encontram os novos caminhos baseadas nas seguintes informações: (1) quantidade de feromônio depositada em cada variável do espaço de busca (elementos espectrais), (2) medida heurística da qualidade de cada caminho (grupos analisados) e (3) um fator randômico de seleção dentro de todo intervalo de busca disponível. A busca das formigas é, portanto, baseada na combinação probabilística entre esses fatores e os possíveis desvios do melhor caminho avaliado (Dorigo, Di Caro *et al.* 2000, Dorigo and Blum 2005).

1.2 Objetivos do Trabalho

Este trabalho teve por propósito estudar e avaliar a viabilidade da aplicação de técnicas de modelagem quimiométrica baseadas em modelos do tipo MISO (Multiplas Entradas Saída Única - *Multiple Input Single Output*) utilizando como dados de entrada componentes espectrais puros, e como variáveis de saída, predições de variáveis de interesse dos processos em análise. Para seleção e combinação dos elementos espectrais individuais, em conjuntos de elementos, será feito uso da técnica de otimização estocástica de colônia de formigas (ACO). A ferramenta formada pela aplicação conjunta de modelagem quimiométrica a partir de grupos de elementos espectrais individuais e do algoritmo ACO foi comparada com resultados de predição de estratégias de quimiometria padrões. Para tanto, dois conjuntos de dados experimentais foram utilizados neste trabalho: O primeiro é composto por dados de amostras de farinhas, caracterizadas por espectroscopia de Infravermelho Próximo, enquanto o segundo é composto por dados de dois ensaios fermentativos da levedura *Saccharomyces cerevisiae* crescendo em meio com glicose e caracterizado com dados de Espectroscopia de Fluorescência 2D.

Objetivos:

- (O1) Desenvolvimento de uma ferramenta integrada para seleção e ajuste de variáveis espectrais puras para inferência de estados;
- (O2) Caracterizar espectroscopia no conjunto amostral de farinha de trigo e centeio, mensurado com Infravermelho Próximo, através da ferramenta de ajuste proposta;
- (O3) Caracterização de fermentação de glicose com a levedura *Saccharomyces cerevisiae*, utilizando a ferramenta proposta e dados de espectroscopia de fluorescência 2D;

- (O4) Comparação entre resultados de predição utilizando a ferramenta proposta e modelos quimiométricos obtidos por PCR e PLS para diferentes conjuntos de dados amostrais;
- (O5) Avaliação de aplicabilidade da ferramenta proposta para pré-seleção de dados de espectro através da comparação com a metodologia de modelagem e filtragem CSMWPLS, em uma versão modificada;
- (O6) Avaliação do incremento na qualidade de predição de modelos PCR e PLS com a utilização de pré-tratamento de dados proposto;
- (O7) Avaliação da viabilidade de caracterização de sistemas fermentativos através de sensores personalizados para este processo.

1.3 Estrutura do Trabalho

Neste capítulo é feita a apresentação geral deste trabalho, bem como os principais objetivos e motivações, além de apresentar considerações relevantes ao entendimento do mesmo.

O Capítulo 2 apresenta uma revisão bibliográfica sucinta sobre o tema quimiometria, descrevendo as técnicas usuais, vantagens e desvantagens, além de introduzir a teoria necessária para o entendimento do procedimento de calibração e validação de modelos multivariados. Também são apresentados os conceitos fundamentais a respeito do algoritmo de otimização estocástica de colônia de formigas e a teoria e conceitos por trás das técnicas de espectroscopia utilizadas como estudos de caso deste trabalho (espectroscopia no infravermelho próximo e fluorescência 2D).

Uma vez que esta tese foi estruturada na forma de artigos científicos, o Capítulo 3 faz um apanhado detalhado sobre os dois conjuntos de dados experimentais utilizados no decorrer do trabalho, necessário para a compreensão completa da problemática abordada nos capítulos seguintes. Também apresenta o detalhamento da ferramenta de modelagem quimiométrica proposta, modelagem Quimiométrica de Elementos Espectrais Puros (PSCM), que combina modelos MISO (Múltiplas Entradas com Saída Única) e seleção de grupos de variáveis espectrais usando ACO (Otimização Colônia de Formigas).

No Capítulo 4 é feita a avaliação da caracterização de amostras de farinha (trigo e centeio) através da utilização de dados de espectroscopia NIR, utilizando a metodologia PSCM. A qualidade de predição obtida por essa metodologia é comparada com metodologias quimiométricas padrões (PCR e PLS). Nesse capítulo também é descrita e avaliada a capacidade da metodologia PSCM para seleção e filtragem de regiões espectrais.

No Capítulo 5, a metodologia PSCM é aplicada na caracterização do processo fermentativo de glicose com a levedura *Saccharomyces cerevisiae*, utilizando dados de espectroscopia por fluorescência 2D, com foco em determinar as melhores regiões espectrais para a predição de cada um dos estados de interesse desse processo, possibilitando o projeto de sensores personalizados, baseados nesta técnica.

No Capítulo 6, a capacidade de seleção e filtragem de dados espectrais inerentes ao PSCM é utilizada em conjunto com o método CSMWPLS, também indicada para

modelagem e seleção de regiões espectrais, de forma a obter uma estratégia híbrida de filtragem e modelagem de processos. A avaliação da eficiência do uso conjunto dessas técnicas é feita com o conjunto de dados experimentais de farinhas de trigo e centeio caracterizadas com medidas de espectroscopia no infravermelho próximo.

O Capítulo 7 apresenta as principais conclusões relativas ao trabalho, além das principais considerações e contribuições. Ainda são apresentadas sugestões para trabalhos futuros, visando a utilização das metodologias desenvolvidas neste trabalho para a proposta de sensores de processos.

1.4 Produção Cientifica

Os diversos capítulos que compõem esta tese serão (ou já foram) publicados em revistas científicas. Além disso, as metodologias desenvolvidas neste trabalho fizeram parte de publicações em congressos, nacionais e internacionais, assim como contribuíram para a realização de trabalhos de conclusão de curso. A seguir são listadas as contribuições técnicas da tese desenvolvida:

Capítulos deste trabalho:

Capítulo 4: RANZAN, STROHM, RANZAN, TRIERWEILER, HITZMANN, and TRIERWEILER (2014). Wheat flour characterization using NIR and spectral filter based on Ant Colony Optimization. *Publicado na revista Chemometrics and Intelligent Laboratory Systems*, Impresso, JCR: 2.291, v. 132, p. 133-140.

Capítulo 5: RANZAN, TRIERWEILER, HITZMANN and TRIERWEILER (2014). Characterization of *Saccharomyces cerevisiae* fermentation using Fluorescence Spectroscopy 2D. *Submetido para a revista Analytica Chimica Acta*.

Capítulo 6: RANZAN, TRIERWEILER, HITZMANN and TRIERWEILER (2014). NIR filtering data using modified Changeable Size Moving Window Partial Least Squares and Pure Spectral Chemometrical Modeling with Ant Colony Optimization for wheat flour characterization. *Submetido para a revista Chemometrics and Intelligent Laboratory Systems*.

Artigos completos publicados em periódicos:

MASIERO, TRIERWEILER, FARENZENA, ESCOBAR, TRIERWEILER e RANZAN (2013). Evaluation of wavelength selection methods for 2D fluorescence spectra applied to bioprocesses characterization. *Publicado na revista Brazilian Journal of Chemical Engineering*, Impresso, JCR: 0.894, v. 30, p. 289-298.

Trabalhos completos publicados em anais de congressos:

• RANZAN, TRIERWEILER, RANZAN, HITZMANN e TRIERWEILER (2012). Caracterização de fermentaçãoes de glicose com *Saccharomyces cerevisiae* com Espectroscopia de Fluorescência 2D utilizando Ant Colony Optimization. Este trabalho foi apresentado no XIX Simpósio Nacional de Fermentações – Sinaferm 2013, em Foz do Iguaçu.

• RANZAN, TRIERWEILER, RANZAN, HITZMANN e TRIERWEILER (2012). Caracterização de fermentaçãoes de glicose com *Saccharomyces cerevisiae* com Espectroscopia de Fluorescência 2D utilizando Ant Colony Optimization. Este trabalho foi apresentado no XIX Simpósio Nacional de Fermentações – Sinaferm 2013, em Foz do Iguaçu.

• RANZAN, TRIERWEILER, HITZMANN e TRIERWEILER (2012). Fluorescence Spectroscopy as a Tool for Ethanol Fermentation On-line Monitoring. *Apresentado no International Symposium on Advanced Control of Chemical Processes - ADCHEM 2012*, em Singapura.

• RANZAN, TRIERWEILER, HITZMANN e TRIERWEILER (2011). Validação com Dados de Espectroscopia de Fluorescência 2D de Modelo Dinâmico para Fermentações Batelada de *Saccharomyces cerevisiae*. *Este trabalho foi apresentado no Simpósio Nacional de Fermentações – Sinaferm 2011*, em Caxias do Sul.

• RANZAN, TRIERWEILER, HITZMANN e TRIERWEILER (2011). Modelagem Quimiométrica para Caracterização de Bioprocessos através de Espectroscopia de Fluorescência -2D. *Este trabalho foi apresentado no Simpósio Nacional de Fermentações – Sinaferm 2011*, em Caxias do Sul.

• ESCOBAR, WALLER MASIERO, RANZAN, TRIERWEILER, FARENZENA e TRIERWEILER (2011). Aplicação de Splines para a Obtenção de Modelos Quimiométricos Voltados à Caracterização de Bioprocessos. *Este trabalho foi apresentado no Simpósio Nacional de Fermentações – Sinaferm 2011*, em Caxias do Sul.

Trabalhos completos submetidos para congressos:

• RANZAN, STROHM, TRIERWEILER, HITZMANN e TRIERWEILER (2014). Modelagem quimiométrica baseada em janela móvel de regressão por mínimos quadráticos parciais associada com otimização de colônia de formigas. *Trabalho submetido para o XX Congresso Brasileiro de Engenharia Química – COBEQ 2014* , em Florianópolis.

• RAMIREZ, RANZAN, FARENZENA e TRIERWEILER (2014). Novo Método para a Quantificação rápida de lipídios em microalgas. *Trabalho submetido para o XX Congresso Brasileiro de Engenharia Química – COBEQ 2014* , em Florianópolis.

• BRESOLIN, RANZAN, RAMIREZ, TRIERWEILER, FARENZENA e TRIERWEILER (2014). Estudo da quantificação de lipídios microalgais através de fluorescência 2D. *Trabalho submetido para o XX Congresso Brasileiro de Engenharia Química – COBEQ 2014* , em Florianópolis.

• RANZAN, RANZAN, TRIERWEILER e TRIERWEILER (2014). Desenvolvimento de Inferências utilizando o Método de Otimização Colônia de Formigas. *Trabalho submetido para o XX Congresso Brasileiro de Engenharia Química – COBEQ 2014* , em Florianópolis.

Contribuição em Trabalhos de Conclusão de Curso:

• AXEL STROHM (2013). Untersuchung Von NIR – und Fluoreszenzspektren zur Charakterisierung von Mehlen. Orientador: HITZMANN, Coorientador: HINRICHS.

Diplomarbeit, Studiengang Lebensmitteltechnologie an der Fakultat Naturwissenschaften - Universität Hohenheim, Stuttgart, Alemanha.

• LUCAS RANZAN (2012). Uso de Ferramentas Matemáticas para Análise de Modelos para Predição de Enxofre e HPA em amostras de Diesel. Orientador: TRIERWEILER, Coorientadores: TRIERWEILER, MUNIZ e RANZAN. *Trabalho de diplomação em Engenharia Química, Departamento de Engenharia Química – UFRGS*.

• LARISSA BRESOLIN (2013). Estudo da quantificação de lipídios microalgais através de fluorescência 2D. Orientador: TRIERWEILER, Coorientadores: TRIERWEILER e RANZAN. *Trabalho de diplomação em Engenharia Química, Departamento de Engenharia Química – UFRGS*.

• CARINE DE VASCONCELOS ALVES (2012). Uma nova sistemática para análise de enxofre em diesel baseada em fluorescência. Orientador: TRIERWEILER, Coorientadores: TRIERWEILER e MUINIZ. *Trabalho de diplomação em Engenharia Química, Departamento de Engenharia Química – UFRGS*.

Contribuição em Trabalhos de Iniciação Cientifica:

• RANZAN (2011). Uso da técnica de Espectroscopia de Fluorescência 2D (EF2D) como ferramenta viável para a caracterização de bioprocessos. Orientador: TRIERWEILER, Coorientadores: TRIERWEILER e RANZAN. *Este trabalho foi apresentado no XXIII Scientific Initiation Meeting - 2011*, UFRGS, em Porto Alegre.

1.5 Contribuições

Pode-se listar como principais contribuições desta Tese os seguintes pontos:

- (C1) ACOW – Função implementada em ambiente Matlab que integra os conceitos de ajuste de modelos MISO com a metodologia de seleção de elementos através do método de otimização de colônia de formigas. Essa ferramenta permite a seleção e ajuste de modelos quimiométricos baseados em grupos de elementos espectrais;
- (C2) PSCM – Metodologia assistida para modelagem quimiométrica e projeto de inferidores de estados de processo a partir de dados de espectroscopia molecular;
- (C3) Mapeamento do plano espectral de fluorescência 2D de fermentações de glicose com a levedura *Saccharomyces cerevisiae*, em função das variáveis de estado deste sistema;
- (C4) PSCM/ACO – Estratégia para seleção e filtragem de dados espectrais baseada em informações qualitativas obtidas utilizando a função ACOW;
- (C5) Estratégia de pré-tratamento de dados e modelagem quimiométrica através da aplicação conjunta de CSMWPLS e PSCM/ACO;
- (C6) modCSMWPLS - Modificação da estratégia CSMWPLS para aplicação em seleção de regiões espectrais;

- (C7) Modelo contínuo de taxas de crescimento de *Saccharomyces cerevisiae* em processos fermentativos de glicose.
- (C8) Ferramenta para projeto de sensores de processo baseados em dados de espectroscopia molecular.

1.6 Resumo Gráfico

A Figura 1.3 apresenta um resumo gráfico que relaciona os objetivos deste trabalho, com os capítulos onde estes objetivos foram tratados, e as respectivas contribuições resultantes desses estudos. A partir desta figura é possível ter uma visão geral de como este trabalho foi estruturado, bem como das interligações entre as diferentes etapas que levaram à conclusão do mesmo.

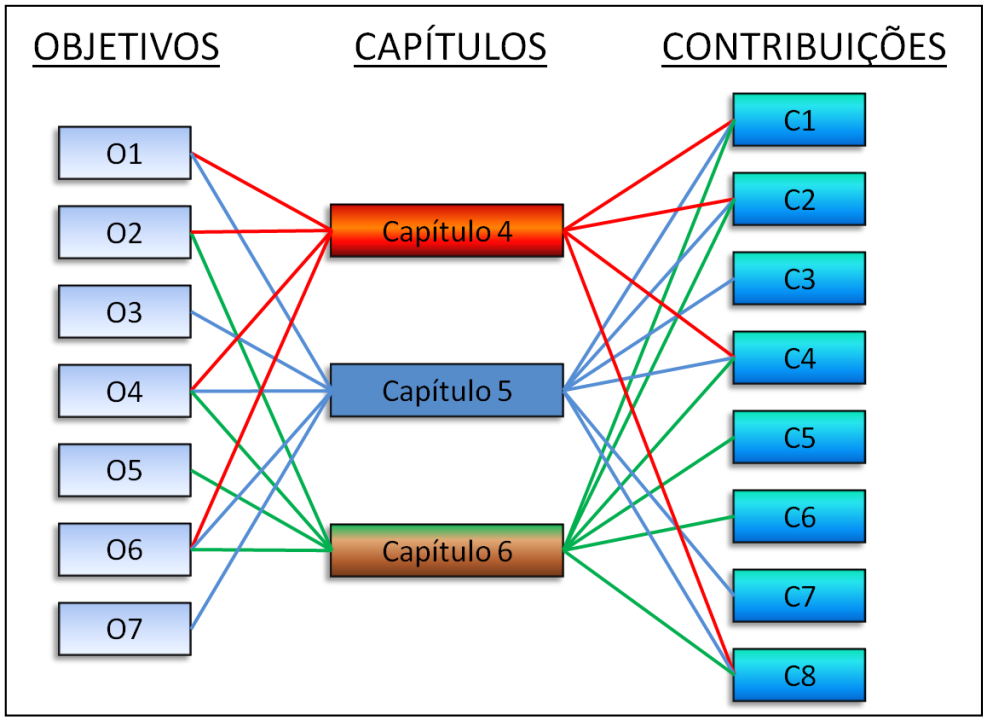

Figura 1.3: Resumo gráfico, indicando as conexões e correlações entre os objetivos, contribuições e capítulos deste trabalho.

Capítulo 2 – Quimiometria e Espectroscopia

2.1 Quimiometria

A quimiometria trata da aplicação de ferramentas matemáticas e estatísticas para o planejamento, otimização e extração de informações partindo de um conjunto de dados físico-químicos multivariados. Introduzida no final dos anos 60, por grupos de pesquisa das áreas de química analítica e físico-química orgânica, seu desenvolvimento foi devido à disponibilidade de equipamentos de análise com respostas multivariadas, assim como a disponibilidade de microprocessadores com elevada capacidade computacional, o que permitiu o desenvolvimento de métricas capazes de tratar grande quantidade de informações de forma simultanea (Geladi 2003).

O surgimento da quimiometria foi resultado do fato das análises estatísticas monovariadas não serem capazes de descrever processos químicos multivariados, obrigando os pesquisadores a desenvolver técnicas que atuassem na análise exploratória e calibração multivariada, de forma a interpretar esta nova estrutura de análises químicas. A quimiometria, portanto, busca encontrar as relações entre variáveis mensuradas e amostras, convertendo essas informações em vetores latentes (Kumar, Bansal *et al.* 2014).

Com relação à análise exploratória, busca-se encontrar quais as variáveis que mais afetam determinado processo, bem como as interações entre elas, de forma a determinar as melhores condições de análise. De maneira geral, esses métodos podem ser classificados como métodos supervisionados, nos quais estão enquadrados (1) Análise discriminante linear (LDA – *Linear Discriminant Analysis*), (2) Método dos K-vizinhos mais próximos (KNN – *K-Nearest Neighbor*), (3) Análise Descriminante com Calibração Multivariável por Mínimos Quadrados Parciais (PLS-DA – *Partial Least Square Discriminant Analysis*), ou não supervisionados como, por exemplo, (1) Análise de Componentes Principais (PCA – *Principal Component Analysis*) e (2) Análise de Agrupamento Hierárquico (HCA – *Hierarchical Cluster Analysis*) (Wehrens 2011).

Os métodos de análise exploratória permitem a interpretação multivariada de conjuntos de dados complexos por meio de gráficos bi- ou tridimensionais. Os métodos supervisionados possuem a desvantagem de necessitar de informações prévias a respeito das amostras, para possibilitar a formação de classes e desenvolvimento de modelos baseados nas informações contidas nas mesmas. Entretanto, nos métodos não supervisionados, a separação de classes ocorre baseada em similaridades estatísticas dos dados espectrais, sem a necessidade de informações sobre a natureza das amostras. Dentre os objetivos dos referidos métodos estão a identificação de grupamentos naturais entre as amostras e o reconhecendo de tendências e padrões (Brereton 2007).

Por fim, a calibração multivariada tem por objetivo estabelecer modelos capazes de relacionar uma elevada quantidade de medidas (variáveis independentes) químicas, físico-químicas ou espectrais de uma dada amostra para inferir os valores de determinadas propriedades de interesse daquela (Kowalski 1983). Dentre os métodos que estão classificados neste grupo, pode ser citada a (1) Regressão Multivariável, (2) Regressão de Componentes Principais (PCR – *Principal Component Regression*), (3) Regressão por Mínimos Quadráticos Parciais (PLSR – *Partial Least Squares Regression*), (4) Regressão de Picos (RR – *Ridge Regression*), dentre outros (Wehrens 2011).

2.1.1 Pré-Tratamento de dados

Na grande maioria dos casos, dados de processo real possuem características marcantes, principalmente quando comparados com dados teóricos; eles são desorganizados, ruidosos, incompletos, deficientes, altamente correlacionados, ou, muitas vezes uma combinação destes. Devido a estas não idealidades na forma como dados de processo são apresentados, a primeira etapa em qualquer análise de dados consiste no pré-processamento para ter acesso aos dados propriamente ditos e possivelmente melhorar a qualidade dos resultados. Esta etapa pode demandar mais tempo do que a análise em si, e mais comumente, este processo consiste em um procedimento iterativo onde as etapas de pré-processamento são intercaladas com as de análise (Beebe, Pell *et al.* 1998).

Alguns problemas podem ser facilmente reconhecidos, assim como ruído de medidas, picos de sensor ou valores anômalos. Neste caso, a escolha da ação ideal não é um problema. Entretanto, a dificuldade surge nas situações em que não se sabe qual característica dos dados contém informação real e quais não. Alguns métodos considerados padrões para pré-tratamento de dados são (1) redução de ruído, (2) ajuste da linha de base, (3) alinhamento de picos, (4) seleção de picos e (5) escalamento (Wehrens 2011).

O método de escalonamento aplicado pode mudar completamente os resultados de uma análise, de forma que a seleção do método (caso exista algum método ideal aplicável ao respectivo conjunto de dados analisado) deve ser cuidadosamente realizado. O escalonamento possui diversos objetivos. Muitos métodos analíticos fornecem dados que não se encontram em escalas absolutas, impossibilitando comparações diretas entre diferentes amostras. Caso algum tipo de padrão interno esteja presente, este pode ser usado para a calibração das intensidades.

Na prática, dados brutos provenientes de análises físico-químicas apresentam variabilidade proveniente de numerosas fontes (p.ex.: variações do meio reacional, ruído, interferências externas, etc). Nesses casos, os pré-tratamentos de dados mais aplicados são a (1) centralizar os dados na média, que consiste na subtração dos elementos de cada variável pela média desta no conjunto amostral, (2) escalonamento, onde cada elemento da variável é dividido pelo desvio padrão da mesma no conjunto amostral, fazendo com que cada variável fique com o mesmo peso e (3) auto-escalonamento, que consiste na centralização dos dados na média e no seu escalonamento (uso simultâneo dos procedimentos 1 e 2) (Wehrens 2011).

Especificamente para dados espectrais, auto-escalonamento não é usualmente recomendado. Muito comumente, os dados consistem de áreas com alta quantidade de informação, contendo picos de diferentes intensidades e áreas contendo apenas ruídos. Quando cada variável espectral é normalizada ao mesmo desvio padrão, o ruído é convertido ao mesmo grau de importância dos sinais que contém a informação atual. Esta situação é claramente indesejável, e nestes casos, a centralização na média é mais aconselhável.

Métodos de pré-processamento especializados são amplamente aplicados em dados de espectroscopia. Quando a intensidade espectral total é dependente da amostra, espectros devem ser escalonados de forma que as intensidades possam ser comparadas. Além dos métodos de escalonamento acima apresentados, o método de normalização usualmente aplicado, especificamente em aplicações com NIR, é o SNV (Standard Normal Variate). Este método essencialmente faz o escalonamento nas amostras ao invés das variáveis. Isto é, cada espectro terá, após o escalonamento, média zero e desvio padrão 1. Isto fornece dados livres de *offsets* e fatores de multiplicação. Obviamente, a suposição de que todo espectro deva ter a mesma média e variância não é sempre válida. Em certos casos, o fato de dados conterem intensidade superior em um espectro comparado aos demais pode conter informações importantes sobre o processo (Wehrens 2011).

2.1.2 Análise Multivariada

Análise de dados multivariados consiste no exame de elevado número de variáveis, de forma a determinar toda a variação da matriz de estudo. O conceito de análise de variação deve ser entendido como a seleção da informação útil contida nessa matriz, descartando ruídos e demais informações que não são úteis ou relativas ao objeto de estudo almejado (Bosque-Sendra, Cuadros-Rodríguez *et al.* 2012).

Regressão Multivariada é amplamente aplicada para resolução de diversos tipos de análise, seja de um ou mais componentes, sendo o método de Mínimos Quadráticos Ordinários (OLS – Ordinary Least Squares) um dos mais utilizados e representativos (Ramos, Beebe *et al.* 1986). Este método procura encontrar o melhor ajuste para um conjunto de dados, através da minimização da soma dos quarados das diferenças entre o valor estimado (modelo) e os dados observados (experimentais). À diferença entre estes dois conjuntos de valores é dado o nome de resíduos.

Para funcionar adequadamente, é necessário que o erro associado às medidas seja de distribuição aleatória, normal e independente. Outra restrição à aplicação do método OLS é referente à estrutura do modelo utilizado. Este método apenas aceita modelos lineares

nos parâmetros, ou seja, as variáveis devem apresentar, além de pesos equivalentes (importâncias), relações lineares entre si. Modelos não-lineares em relação aos parâmetros necessitam de métodos de regressão não-lineares para sua resolução (Tellinghuisen 2010).

Partindo do pressuposto que o objetivo do método OLS é a determinação do modelo linear capaz de converter a matriz de dados mensurados (***X***) na matriz de variáveis preditas (***Y***) é necessário determinar os valores adequados dos parâmetros do modelo (***b***). Dessa forma, as equações observacionais podem ser expressas, para o caso generalizado multivariável, na forma da equação 2.1 (Nascimento, Froes *et al.* 2010).

$$\boldsymbol{y = b_0 + x_1 b_1 + x_2 b_2 + \cdots + x_k b_k + e} \tag{2.1}$$

O modelo presente na equação 2.1 pode ser expandido para sua forma matricial de acordo com a equação 2.2, ao se considerar *k* variáveis explicativas (independentes) e *n* observações (amostragens).

$$\begin{pmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \\ \vdots \\ \boldsymbol{y_n} \end{pmatrix} = \begin{pmatrix} \boldsymbol{1} & \boldsymbol{x_{11}} & \boldsymbol{x_{21}} & \cdots & \boldsymbol{x_{k1}} \\ \boldsymbol{1} & \boldsymbol{x_{12}} & \boldsymbol{x_{22}} & & \boldsymbol{x_{k2}} \\ & \vdots & & \ddots & \vdots \\ \boldsymbol{1} & \boldsymbol{x_{1n}} & \boldsymbol{x_{2n}} & \cdots & \boldsymbol{x_{kn}} \end{pmatrix} * \begin{pmatrix} \boldsymbol{b_0} \\ \boldsymbol{b_1} \\ \vdots \\ \boldsymbol{b_k} \end{pmatrix} + \begin{pmatrix} \boldsymbol{e_1} \\ \boldsymbol{e_2} \\ \vdots \\ \boldsymbol{e_n} \end{pmatrix} \tag{2.2}$$

A equação 2.2 pode ser escrita em sua forma resumida de acordo com a equação 2.3.

$$\boldsymbol{y = Xb + e} \tag{2.3}$$

A resolução do problema reside na determinação do vetor de parâmetros do modelo (***b***), alcançada através da minimização da soma do quadrado dos resíduos ($\boldsymbol{min\left(\sum_{i=1}^{n} e_i^2\right) = min(e^T e)}$). Após fazer as devidas substituições, é possível chegar à resolução analítica do problema em questão, onde o vetor ***b*** pode ser calculado de acordo com a equação 2.4, e a matriz de variância-covariância dos parâmetros é calculada de acordo com a equação 2.5. O termo σ^2 é chamado de variância residual, sendo tipicamente estimado pela equação 2.6 (Ferraty and Romain 2011).

$$\boldsymbol{b} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T y} \tag{2.4}$$

$$\boldsymbol{Var(b)} = (\boldsymbol{X^T X})^{-1} \boldsymbol{\sigma^2} \tag{2.5}$$

$$\boldsymbol{\hat{\sigma}^2} = \frac{\boldsymbol{1}}{\boldsymbol{n-k-1}} \boldsymbol{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{2.6}$$

O desvio padrão individual dos coeficientes, fornecido pela raiz quadrada dos elementos diagonais da matriz de variância-covariância (equação 2.5), pode ser utilizado para teste estatístico. Variáveis que apresentarem coeficientes não significativamente diferentes de zero são usualmente removidas do modelo.

O OLS possui algumas desvantagens, sendo a mais significativa para o ramo das ciências naturais, a sensibilidade deste método à colinearidade, que é a existência de, pelo menos, uma dependência linear entre as variáveis independentes (Mardia, Kent *et al*. 1979). O intervalo de confiança para os coeficientes de regressão são baseados no pressuposto de independência das variáveis e, por consequência, de parâmetros. Correlações entre variáveis violam esta independência, inviabilizando o cálculo do

conjunto de parâmetros ideal, gerando modelos instáveis, no sentido da variância dos parâmetros estimados ser alta e as predições menos confiáveis. De forma genérica, OLS não permite ajustes em casos onde $(\boldsymbol{X}^T\boldsymbol{X})$ for singular, ou então nos casos em que o inverso não puder ser calculado, indicando situações em que o número de variáveis independentes é maior que o número de observações (Wehrens 2011).

2.1.3 PCA – Análise de Componentes Principais

PCA é uma ferramenta quimiométrica que busca evidenciar similaridades ou diferenças em amostras de um determinado conjunto de dados. De fato, este método define novas variáveis, obtidas através de combinações lineares das originais, de forma que a direção do primeiro novo eixo contém a maior variação. Cada nova variável subsequente é ortogonal às variáveis antecedentes, mas novamente na direção que contém a maior variação remanescente. Desta maneira, o resultado da PCA é um conjunto de novos vetores, ortogonais entre si, chamados de componentes principais (PC – *Principal Component*), resultante do agrupamento de variáveis altamente correlacionadas (Beebe, Pell *et al.* 1998). De forma prática, PCA reduz a dimensionalidade do conjunto de dados original, fazendo com que as informações mais relevantes fiquem concentradas nos primeiros componentes principais (Reddy 2011).

A ideia central por trás da PCA é que, usualmente, dados com grandes dimensões são compostos por muitas variáveis supérfluas. Em uma análise mais detalhada em espectros de alta resolução, conclui-se que comprimentos de onda nas vizinhanças próximas são altamente correlacionados e contêm informações similares. Uma forma de filtragem desses dados poderia ser feita com base na escolha de comprimentos de onda que possuem maior informação, ou então daqueles que diferem dos demais. Este processo pode ser baseado em agrupamento de variáveis e seleção de um representante de cada grupo. Entretanto, referida abordagem é um tanto elaborada e leva a diferentes resultados dependendo dos critérios de agrupamento e descarte empregados.

PCA fornece um mapeamento em poucas dimensões, oriundo de matrizes de dados de grandes dimensões, concentrando a maior parte da informação contida nos dados originais. As coordenadas das amostras no novo espaço são chamadas de *Scores*. As novas dimensões são combinações lineares das variáveis originais e são chamadas de *Loadings*. O termo componente principal pode se referir a ambas, *Scores* e *Loadings*, dependendo do contexto empregado. A multiplicação matricial dos *Scores* com a matriz transposta de *Loadings* leva a uma aproximação da matriz original de dados. É importante salientar que o número máximo de PCs (componentes principais) de uma dada matriz de dados é igual a menor dimensão desta matriz, seja o número de amostras ou variáveis (Jolliffe 1986).

Os PCs são combinações ortogonais das variáveis, definidos de forma que a variância dos *Scores* é máxima, a soma Euclidiana das distâncias ente os *Scores* é máxima e a reconstituição da matriz de dados original ($\widetilde{\boldsymbol{X}}$) é o mais próxima possível da matriz original $\boldsymbol{X}$ ($\left\|\boldsymbol{X}-\widetilde{\boldsymbol{X}}\right\|$ é mínimo) (Jackson 1991).

A técnica de PCA possui inúmeras vantagens: é simples, possui uma única solução analítica e geralmente leva a uma representação dos dados mais simples de ser interpretada. A desvantagem desse método é que ele não produz, como resultado, um

pequeno grupo de comprimentos de onda que carregam consigo a informação, mas sim um pequeno grupo de PCs, nos quais todos os comprimentos de onda estão representados.

Uma vez que a PCA definiu as variáveis latentes, todas as amostras podem ser graficadas, ignorando os PCs de ordem maiores. Usualmente, poucos PCs são necessários para capturar a maior fração de variância do conjunto de dados (apesar disto ser altamente dependente da característica dos dados analisados) (Geladi 2003).

A implementação da PCA padrão pode ser feita até mesmo em softwares com baixo poder numérico, isso porque o algoritmo utilizado para o cálculo dos PCs é a Decomposição em Valores Singulares (SVD – *Singular Value Decomposition*) da matriz de dados. Uma alternativa à utilização de SVD seria a decomposição da matriz de covariância ou correlação dos dados em seus autovalores e autovetores, entretanto, SVD é numericamente mais estável e preferível na grande maioria dos casos.

SVD decompõe uma matriz de dados ***X***, de dimensões *n x p*, centralizada na média, em três partes, de acordo com a equação 2.7. Nessa equação, ***U*** é uma matriz ortonormal de dimensão *n x a* que contêm os vetores singulares à esquerda. ***D*** é a matriz diagonal (*a x a*) que contém os valores singulares, e ***V*** é a matriz ortonormal de dimensão *p x a* que contêm os vetores singulares à direita, também chamados de *loadings*. Os resultados da multiplicação das duas primeiras matrizes (***U*** e ***D***) são chamados de *scores* na terminologia da PCA. A variável *a* indica o número de componentes principais, sendo equivalente ao menor número entre linhas e colunas da matriz ***X*** ($\boldsymbol{a_{max} = \min(n, p)}$).

$$\boldsymbol{X = (UD)V^T = TP^T} \tag{2.7}$$

A interpretação das matrizes ***T***, ***P***, ***U***, ***D*** e ***V*** é direta. Os *loadings*, colunas da matriz ***P*** (ou equivalentemente, os vetores singulares à direta, colunas na matriz ***V***) fornecem os pesos das variáveis originais nos PCs. Variáveis que possuem valores muito pequenos em colunas específicas de ***V*** possuem contribuição ínfima com aquela variável latente particular. Os *scores*, colunas em ***T***, constituem as coordenadas no espaço de variáveis latentes; estas são as coordenadas das amostras no novo espaço da PCA. As colunas em ***U*** fornecem as mesmas coordenadas em uma forma normalizada (variância unitária), enquanto as colunas em ***T*** possuem variância correspondente as variâncias equivalentes à cada PC. Estas variâncias, λ_i, são proporcionais aos quadrados dos elementos diagonais da matriz ***D***, como mostra a equação 2.8, onde d_i representa o elemento *i,* diagonal da matriz ***D***.

$$\boldsymbol{\lambda_i = d_i^2/(n-1)} \tag{2.8}$$

A variância explicada (***FV***) de cada PC pode ser obtida de acordo com a equação 2.9 (Camacho, Picó *et al.* 2010).

$$\boldsymbol{FV(i) = \lambda_i / \sum_{j=1}^{a} \lambda_j} \tag{2.9}$$

Apesar da simplicidade de implementação desse método, o principal problema da aplicação da PCA recai sobre a dificuldade de escolha do número ideal de PCs a serem utilizados na descrição de um determinado conjunto de dados. Em outras palavras, determinar quando termina a informação e começa o ruído. Existem métodos que avaliam a quantidade de variância explicada, ou então utilizam testes gráficos e

estatísticos para definir quais PCs devem ser utilizados. Entretanto, mesmo esta escolha fica a critério do pesquisador.

A quantidade de variância de cada PC é usualmente retratada na forma do gráfico de Scree (*Scree Plot*), onde as próprias variâncias, ou alternativamente, os logaritmos destas, são apresentados na forma de barras. Usualmente é considerado a fração de variância explicada total para cada PC individualmente. Os últimos PCs, na maioria dos casos, não contêm informação, e especialmente na escala logarítmica, tendem a fazer o gráfico de Scree menos interpretativo, portanto são normalmente descartados desta avaliação. A Figura 2.1 apresenta um exemplo de análise de PCs através do gráfico de Scree, adaptado de Wehrens (2011), para um conjunto de dados genéricos.

Os gráficos contidos na Figura 2.1 claramente mostram que para o referido exemplo, os PCs 1 e 2 possuem mais variância que os demais. Juntos, eles possuem 55% do total de variância acumulada. Apesar disto, os gráficos de Scree não apresentam um ponto de corte claro (fato não usualmente apresentado por dados reais). Dependendo do objetivo da investigação, poderiam ser escolhidos de três à cinco PCs. A escolha de quatro PCs, neste caso, não faria sentido, pois o quinto PC iria adicionar tanta variância quanto o quarto PC, de forma que se o quarto PC for adicionado, assim deve ser feito com o quinto.

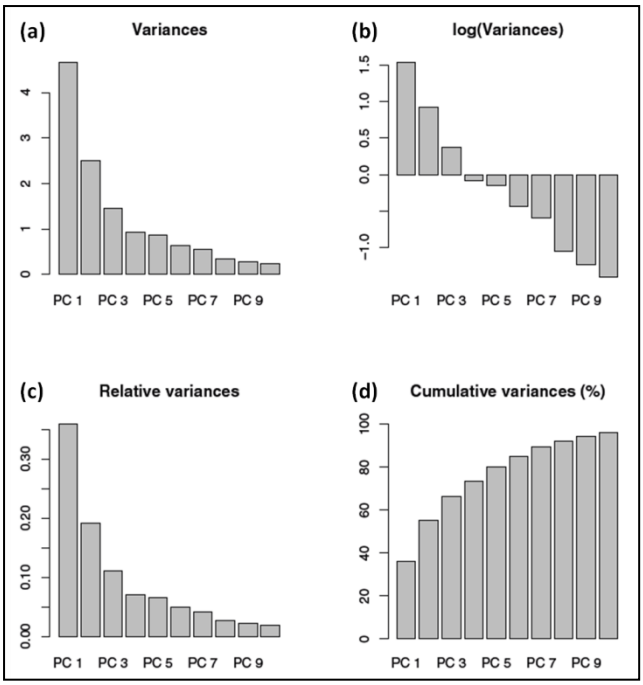

Figura 2.1: Gráficos de Scree para avaliação do total de variância explicada para cada PC. (a) Variâncias para cada PC, (b) Logaritmo das Variâncias, (c) Fração do total de variância acumulada e (d) percentagem cumulativa do total de variância. **Fonte: Adaptado de Wehrens (2011).**

Uma vez que determinado conjunto de dados foi redefinido para um número de dimensões menor por análise de PCA, é possível determinar como outros dados serão posicionados neste novo espaço dimensional. Estes novos dados correspondem a novas amostras, medidas ou instrumentos. O objetivo é que a representação em espaços dimensionais menores permita uma avaliação dos dados de forma mais ampla e simples para a identificação de padrões, possível através da obtenção dos *Scores* do novo

conjunto de dados. Partindo de uma nova matriz de dados ***X***, a projeção desta no espaço definido pelos *Loadings (**P**)*, pode ser obtida de acordo com a equação 2.10.

$$\boldsymbol{XP} = \boldsymbol{TP^T P} = \boldsymbol{T} \tag{2.10}$$

Um detalhe importante é que o escalonamento da nova matriz de dados deve ser feita com as médias e desvios padrões utilizados na análise de PCA inicial, uma vez que ambos os conjuntos de dados devem possuir o mesmo ponto de origem.

2.1.4 PCR – Regressão de Componentes Principais

PCR é método de regressão baseado em fatores que resolve algumas limitações do método OLS. Este método expressa a informação contida em diversas variáveis medidas através de um novo sistema de coordenadas composto por apenas alguns componentes.

A ideia primordial do PCR é utilizar os *Scores* obtidos na PCA, ao invés do conjunto de dados original, no passo de regressão para obtenção do modelo preditivo de uma determinada variável de interesse. Este método apresenta duas grandes vantagens em relação à aplicação da técnica de OLR aos dados originais: (i) como os *Scores* são ortogonais, não há problemas com variáveis correlacionadas na matriz de variáveis independentes, de forma que a inversão da matriz de covariância é simples; e (ii) o número de PCs levados em consideração é significativamente menor que o número de variáveis originais, diminuindo consideravelmente o montante de parâmetros a serem estimados, levando a um maior número de graus de liberdade para a estimação dos erros.

O preço pago pelas vantagens de utilização PCR recai sobre a possível perda de informações vitais do processo, devido à compressão de dados, a necessidade de determinar o grau de compressão utilizado (seleção do número de componentes principais utilizados) e por fim, a incapacidade de obtenção de expressões analíticas para predição do erro e das variâncias dos coeficientes de regressão individuais. Conclusões acerca do número ótimo de componentes principais e erro esperado de predição só podem ser feitas através da utilização de técnicas de Validação Cruzada (*Cross validation*) ou similares.

A metodologia PCR pode ser dividida em duas etapas distintas: Obtenção dos componentes principais a partir da matriz de dados original (aplicação da PCA) e regressão utilizando estes componentes para obtenção do modelo de predição. Uma vez obtidos os *Scores* e selecionados os PCs mais importantes que serão utilizados na regressão, o problema fica resumido à resolução do problema de regressão do modelo, resumido na forma da equação 2.11, onde ***T*** é uma matriz *n* x *a*, onde *n* corresponde ao número de observações, *a* ao número de componentes principais selecionados, ***Y*** corresponde as observações da variável de interesse e ***A*** é o vetor de parâmetros do modelo.

$$\boldsymbol{A} = (\boldsymbol{T^T T})^{-1} \boldsymbol{T^T Y} \tag{2.11}$$

Na prática, PCR é aplicado em matrizes de dados centralizados na média. Sem este tratamento nos dados, o primeiro PC, usualmente, apresenta variância significativamente elevada em comparação aos demais, e valores muito próximos ao vetor de média das colunas da matriz de dados original. Outro cuidado que deve ser tomado é com relação

ao auto-escalonamento, pois este procedimento pode salientar o ruído de amostras não significativas com relação ao analito avaliado (Geladi 2003, Wehrens 2011, Hemmateenejad, Miri *et al.* 2012, Kumar, Bansal *et al.* 2014).

2.1.5 PLSR – Regressão por Mínimos Quadrados Parciais

No PCR, a informação contida nas variáveis independentes é resumida em um pequeno número de componentes principais, entretanto, não existe nenhuma garantia de que os PCs associados aos maiores valores singulares sejam os mais indicados para a regressão. Mesmo sendo o PC 1 aquele que possui a maior variância explicada, pode ser que ele ainda possua capacidade de predição limitada. Uma vez que a escolha dos PCs é feita de forma sequencial partindo do primeiro, existe a chance real de serem inclusas variáveis que não acrescentem informações significativas ao modelo ajustado. Resumidamente: a informação contida na matriz de dados original é comprimida sem nenhum tipo de consideração com relação ao que será predito, de forma que não pode ser garantido que a informação essencial contida nos dados será preservada (Krishnan, Williams *et al.* 2011, Cariou, Verdun *et al.* 2014).

PLSR, assim como o PCR, define variáveis latentes ortogonais para comprimir a informação e descartar dados irrelevantes. Entretanto, PLS visa obter as variáveis latentes de forma a capturar a maior variância contida em $\boldsymbol{X}$ e $\boldsymbol{Y}$, maximizando a correlação entre estas matrizes (Dayal and MacGregor 1997).

O PLSR trata-se de um método iterativo, onde seus componentes são extraídos sequencialmente, diferentemente do PCR, onde todos os PCs são determinados simultaneamente em uma única decomposição SVD. Em cada iteração, a variação associada com o componente estimado é retirada dos dados, em um processo chamado redução (*deflation*), enquanto a variação remanescente é utilizada para a estimação do próximo componente. Este processo é repetido enquanto o usuário desejar, ou então até que todos os vetores de carga tenham sido calculados.

O primeiro componente é obtido a partir do SVD da matriz de produto cruzado, $\boldsymbol{S = X^T Y}$, onde o primeiro vetor singular à esquerda, *w*, pode ser interpretado como a direção de variância máxima da matriz de produto cruzado, e é usualmente indicado com a descrição de "pesos". A projeção da matriz $\boldsymbol{X}$ nesse vetor é chamada de "*X scores" (**t**)*. Estes *Scores* serão reunidos em uma matriz $\boldsymbol{T}$ que desempenha o mesmo papel que a matriz de *Scores* no PCR, apresentando baixa dimensionalidade e estimação da informação contida em $\boldsymbol{X}$.

No processo de regressão de $\boldsymbol{Y}$ utilizando $\boldsymbol{T}$ devem ser obtidos vetores de carga para $\boldsymbol{X}$ e $\boldsymbol{Y}$, através de regressões com o mesmo vetor de *Score,* $\boldsymbol{t}$, o que leva, por fim, à redução das matrizes de dados do estado atual. A partir deste ponto, a estimação do próximo componente pode ter início com a decomposição SVD da matriz de produto cruzado dos dados residuais, ou então o algoritmo é encerrado.

De fato, existem diversos estruturas de algoritmos para o cálculo dos vetores de cargas das matrizes $\boldsymbol{X}$ e $\boldsymbol{Y}$, entretanto, suas diferenças são causadas por variações em critérios de otimização, diferentes implementações para obtenção de melhoras de desempenho e velocidade, ou na escolha de diferentes escalonamentos para resultados intermediários.

De posse dos vetores de *Scores*, obtidos pelo método PLS, a resolução do PLSR encontra-se em uma etapa similar ao PCR, onde ao invés de modelarmos ***Y*** a partir de ***X***, será utilizado os vetores de *Scores (**T***) para calcular os coeficientes de regressão (***A***) do modelo de predição de ***Y***. O vetor ***A*** é calculado de acordo com a equação 2.11, equivalentemente à utilizado no modelo PCR, sendo a única diferença o método aplicado para a obtenção de ***T***, que no PLSR leva em consideração a informação de ***Y*** (Kettaneh, Berglund *et al.* 2005).

PLS apresenta algumas vantagens que motivam sua vasta aplicação nas áreas de modelagem quimiométrica de processos, dentre as quais podemos citar o fato deste método ser considerado resistente ao chamado *overfitting*, que é o termo utilizado para descrever situações onde modelos estatísticos se ajustam em demasiado ao conjunto de amostras usado em sua calibração, se ajustando a erros de medição e fatores aleatórios presentes neste conjunto, mas não representativos do processo como um todo (Land Jr, H. *et al.* 2011)

2.1.6 Seleção de Variáveis – Métodos de Otimização Global

Seleção de variáveis é um tópico muito importante em diversos tipos de modelagem, ao ponto de determinar a eficiência do resultado. Todos os métodos mencionados anteriormente possuem sua capacidade de predição vinculada à seleção de variáveis independentes. Nas abordagens sem supervisão, variáveis pouco informativas podem mascarar o processo “real”. Em casos supervisionados (procedimentos de classificação e regressão juntos), existe o perigo de variáveis pouco correlacionadas às variáveis dependentes levarem a modelos com baixo poder preditivo.

O objetivo da seleção de variáveis é, portanto, reduzir o número de variáveis independentes selecionadas para aquelas que contêm somente informação relevante, isto é aumentar a seletividade, consequentemente, melhorando a modelagem estatística. Esta seleção deve ser analisada em termos de desempenho preditivo (reduzindo o número de chances de correlação entre variáveis) e interpretabilidade (modelos trazem informação a respeito dos sistemas sendo estudados, sendo modelos pequenos mais facilmente interpretados do que aqueles de tamanho elevado). O exemplo mais simples de seleção de variáveis é a seleção de picos em dados de espectroscopia, onde diversos analitos podem ser quantificados utilizando combinações entre elementos espectrais que apresentam picos.

A prática mais comum na seleção de variáveis baseia-se nos próprios coeficientes de ajustes obtidos em regressões usando dados espectrais, onde elementos que apresentam valores mais elevados são mantidos e aqueles cujo valor pode ser considerado estatisticamente insignificante são desprezados. Na etapa seguinte, os elementos previamente desprezados são acrescentados ao modelo e um novo ajuste é realizado, mantendo os elementos com coeficientes mais elevados. Este procedimento é repetido até que o número de variáveis desejado seja alcançado. Infelizmente, coeficientes de elementos altamente correlacionados assumem valores elevados, fazendo com que a ordem de grandeza dos coeficientes não seja um bom indicador de importância.

Usualmente, considera-se que o único modo efetivo de garantir a capacidade de modelagem de um pequeno grupo de variáveis é através do seu teste efetivo. No caso do modelo não apresentar resultados satisfatórios, a solução seria testar um novo grupo de variáveis. Assumindo alguma forma de quantificar o erro de medição, muitos

pesquisadores vêm aplicando algoritmos de otimização na procura dos grupos de variáveis que fornecem a capacidade preditiva máxima para os referidos sistemas (Allegrini e Olivieri, 2011; Kumar *et al.*, 2014; Leardi, *et al.,* 2002; Hemmateenejad, *et al.,* 2011).

A elevada velocidade computacional das máquinas modernas é uma das principais responsáveis pela utilização de métodos de seleção de variáveis dentro de grandes conjuntos de dados, entretanto, mesmo casos com quantidade de variáveis moderadas não permitem a análise de todos os possíveis grupos formados por combinações de seus elementos. A limitação de regiões de busca espectral para tamanhos aceitáveis, como em torno das melhores soluções atuais (método *Stepwise*), pode viabilizar a busca exaustiva, entretanto, limita a avaliação das áreas espectrais, podendo desprezar zonas significativas, fornecendo, na grande maioria dos casos de grandes conjuntos de dados, soluções de ótimos locais com reduzida qualidade preditiva (K. and P. 2009).

Uma alternativa para resolver o problema de regiões desprezadas é a utilização de busca randômica, onde são feitos testes com diferentes conjuntos de variáveis escolhidas de forma aleatória. O problema deste método é que as chances de se encontrar o ótimo global são praticamente nulas. A solução, portanto, é a utilização conjunta de elementos randômicos e informação gradual, através de algoritmos de otimização global que combinem a informação contida em soluções de qualidade elevada, com a habilidade de descartar esta informação caso necessário, possibilitando o algoritmo de sair dos arredores de ótimos locais. Os dois algoritmos de otimização heurística mais conhecidos e aplicados no ramo da quimiometria são o de Recozimento Simulado e o Algoritmo Genético (Kirkpatrick, Gelatt *et al.* 1983, Cerny 1985), entretanto, outras metodologias como a Busca Tabu, Colônias artificiais de abelhas, Enxame de partículas e Busca Harmônica, também podem ser empregadas para esta aplicação (Mello and Pinto 2008, Ghasemi *et al*. 2012).

Duas estratégias distintas são empregadas na seleção de variáveis usando algoritmos de otimização global. A primeira, fixa o tamanho do grupo de variáveis selecionadas, usualmente escolhido baseado em considerações prévias como PCA, e seleciona o grupo de variáveis que apresenta o melhor desempenho. A segunda estratégia penaliza a inclusão de variáveis extras, deixando o algoritmo de otimização determinar, eventualmente, o tamanho ideal do modelo, juntamente com o grupo de variáveis mais indicado.

Dentre os métodos de otimização global conhecidos, os mais amplamente utilizados na seleção de variáveis são os do tipo Algoritmos Genéticos. A ideia principal por trás destes métodos é tratar uma população de possíveis soluções como vetores (cromossomos) formados por valores binários (genes), e através da simulação do que seria a reprodução sexual, mesclar os melhores indivíduos, desta forma, os indivíduos resultantes são formados por combinações das soluções contidas em seus pais. A qualidade dos descendentes é medida em uma etapa de avaliação, chamada de Aptidão (*Fitness*). Vetores com baixas aptidões não terão, ou então terão baixa probabilidade de reprodução, de forma que gerações subsequentes apresentarão cada vez melhores soluções. Este método obviamente imita o processo de seleção natural, onde apenas os indivíduos melhor adaptados têm a chance de se reproduzir, dando origem ao seu nome (Goldberg 1989).

A Busca Tabu, por sua vez, remonta sua origem na década de 70. Consiste em um método iterativo de procura de ótimos globais não monotônico, cuja principal característica é a capacidade de exploração do histórico do processo de busca, organizado em estruturas que compõe a chamada “memória adaptativa”. Este método mantêm uma lista de movimentos proibitivos, conhecida como Lista Tabu, reduzindo o risco de ciclagem (execução em ciclo infinito) do algoritmo (Pereira 2007).

Colônias artificiais de Abelhas são métodos de otimização, inicialmente propostos por Karaboga no ano de 2005, nos quais a busca pelo ótimo global é baseada em informações como a direção de cada solução obtida ate o momento (Silva 2011).

Otimizadores do tipo Enxame de Partículas (PSO) são inspirados na dinâmica comportamental de pássaros, insetos e peixes. Inicialmente propostos para resolução de problemas contínuos, estão sendo aplicados nas mais diversas áreas. O sistema é inicializado o com uma população de soluções aleatórias e procura por um resultado ótimo melhorando gerações. Entretanto, diferentemente de algoritmos genéticos, o PSO não tem operadores de evolução como crossover e mutação. Em PSO as soluções potenciais, chamadas de partículas, voam através do espaço do problema seguindo as então melhores partículas (que possuem os melhores valores no momento). O PSO se baseia na informação da trajetória das partículas (indivíduos) e dos pontos do espaço de busca visitados por elas para informar a qualidade da solução (qualidade da função objetivo). Para tanto, usa-se uma estrutura de memória para preservar os melhores locais visitados. A indicação da movimentação de cada partícula a cada nova iteração depende de duas informações: a melhor posição de todo o enxame e a melhor posição da própria partícula. São introduzidas regras estocásticas no PSO dando o caráter aleatório ao algoritmo (Lacerda 2007).

Recentemente, a técnica de Otimização Colônia de Formigas (ACO – *Ant Colony Optimization*) tem sido aplicada para resolução de problemas envolvendo seleção de variáveis PLSR. ACO é um algoritmo de otimização que foi concebido baseado no comportamento coletivo de formigas quando saem em busca de fontes de alimento. As formigas em um formigueiro executam tarefas coletivas e tomam decisões que requerem um alto grau de coordenação, tais como a construção do formigueiro, alimentação da ninhada, armazenamento e busca de comida e assim por diante (Allegrini and Olivieri 2011).

A Figura 2.2 representa um experimento onde as formigas “descobrem” de forma coletiva qual o menor caminho entre o formigueiro e a fonte de alimentos. Em experimentos conduzidos em laboratório por Goss *et al.* (1989) com a formiga *Linepithaema humilis* uma fonte de alimentos é separada do formigueiro por uma trilha com diferentes caminhos possíveis. Inicialmente as formigas ocupavam todos os possíveis caminhos entre o ninho e a fonte de comida, já, ao final do experimento, todas as formigas ocupavam somente o menor caminho entre os dois pontos. Foi constatado que isto ocorre devido ao fato das formigas, quando em movimento, secretarem no solo certa quantidade de feromônio, marcando seu trajeto.

O feromônio, por ser uma substância biologicamente ativa, que evapora com o tempo, funciona como um chamariz para as demais formigas. No experimento real, as formigas que percorreram o menor caminho, retornaram ao formigueiro mais rapidamente, de forma que a trilha percorrida por estes indivíduos apresenta maior

concentração de feromônio. Novas formigas que se encontrarem em bifurcações com esta trilha ótima, darão preferência à trilha com maior quantidade de feromônio (caminho mais curto e utilizado mais vezes), otimizando a distância percorrida entre o ninho e a fonte de comida.

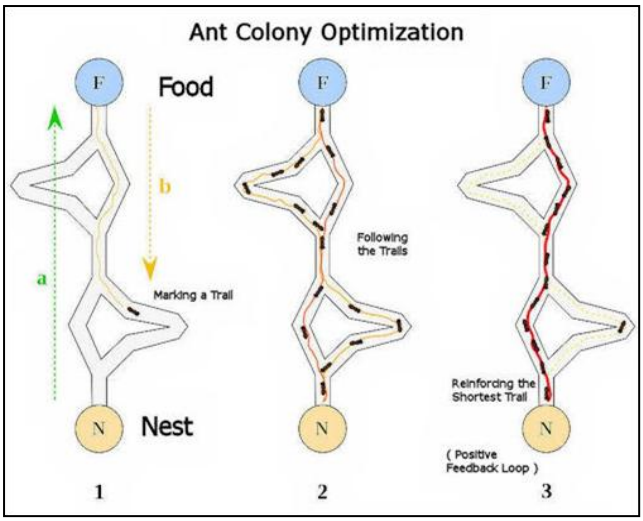

Figura 2.2: Diagrama mostrando a evolução, no tempo, do experimento de Goss *et al.*, (1989) sobre o estudo do processo de busca de comida pelas formigas da espécie *L. humilis*, onde o caminho entre o ninho e a fonte de comida é otimizado através do trabalho conjunto da colônia. **Fonte: Adaptado de Goss *et al.*, (1989)**.

A primeira versão do algoritmo ACO foi desenvolvida por Dorigo e Gambardela (1997), para a resolução do problema do Caixeiro Viajante, um problema de otimização combinatória de busca em um espaço de permutações. Este algoritmo é fundamentado na distribuição de um exército de formigas que deve visitar um conjunto de cidades, uma única vez, percorrendo a menor distância possível (Shamsipur, Zare-Shahabadi *et al.* 2006).

Atualmente diversos trabalhos vêm sendo publicados com relação à aplicação do método ACO na seleção de variáveis, dentre os quais podemos citar os trabalhos de Allegrini e Olivieri, (2011), Dorigo e Blum (2005), Hemmateenejad *et al.* (2011), Mullen *et al.* (2009), Shamsipur *et al.* (2006), Shamsipur *et al.* (2009) e Socha *et al.* (2008).

2.2 Validação de Modelos

Validação de modelos trata da avaliação da qualidade preditiva de modelos. Um modelo capaz de predições acuradas é considerado uma boa descrição da realidade. Validação, entretanto, não é uma tarefa simples. Sistemas químicos ou bioquímicos normalmente apresentam variações significativas, não relacionadas aos parâmetros dos modelos.

Um segundo aspecto relacionado à validação de modelos é referente à quantificação da estabilidade dos coeficientes destes. Em modelos de regressão com dados de espectroscopia é esperado que comprimentos de onda com coeficientes de regressão elevados correspondam aos picos nos espectros, já que na grande maioria dos métodos de espectroscopia é possível associar características espectrais a fenômenos físico-químicos (vibrações específicas, transições eletrônicas, átomos, etc.) (Brereton 2013).

Um aspecto chave é o fato do erro de estimação, assim como o intervalo de confiança dos modelos, ser diretamente derivado do conjunto de dados disponível para treinamento/calibração dos modelos, de forma que os modelos somente serão representativos do processo em estudo se o conjunto de amostras também o for. Caso existam diferenças sistemáticas entre o conjunto de amostras utilizado para calibração e o conjunto de amostras destinada à predição, as predições não serão ótimas, correndo o risco de serem totalmente erradas. Tais diferenças podem ser fruto de diversos aspectos: diferentes operadores, fornecedores de soluções químicas ou equipamentos, métodos de coleta, etc. Estes fatores podem fazer com que os dados de calibração e teste sejam levemente, mas consistentemente, diferentes, o que pode inviabilizar a utilização dos modelos.

Especialmente em grandes conjuntos de dados, validação é normalmente baseada em uma única divisão entre conjunto de dados de calibração e conjunto de dados de teste. Caso o número de amostras seja suficientemente elevado, isso evitará o fenômeno de *overfitting* e os erros de estimação correspondentes serão reduzidos, dependendo apenas de como os conjuntos de calibração e teste são divididos.

A divisão randômica é aconselhável para a maioria dos casos, sendo necessária a conferência da representatividade do conjunto de dados de calibração, de forma a evitar a seleção de conjuntos de amostras com características especificas e distintas do grupo geral da população de amostras. Esta verificação pode ser feita através da análise de PCA do grupo amostral total. O grupo de calibração deve ser, não apenas representativo do grupo de teste, mas também independente (Kowalski 1983).

Uma vez que o grupo de amostras disponível pode ser considerado representativo do conjunto de dados que serão coletados e preditos, deve ser adotada uma forma de estimar a qualidade das predições, simulando através dos dados coletados, dados que serão preditos.

Na validação cruzada (*Cross Validation*) a simulação de dados futuros é feita através da retirada de parte das amostras coletadas e treinando os modelos usando o conjunto de amostras remanescentes. Este é um método simples e confiável para estimação de erros de predição.

Na validação cruzada, assim como em outros métodos de avaliação de modelos, são necessários parâmetros métricos para quantificação da qualidade preditiva dos modelos, dentre os quais a raiz quadrada do erro médio de predição (RMSEP – *Root-Mean-Square Error of Prediction*) e o coeficiente de determinação R^2 são os mais utilizados. Além destes, a raiz quadrada do erro médio de calibração (RMSEC) é a variação do RMSEP aplicada ao conjunto de dados do grupo de calibração, sendo crucial para a aplicação de rotinas de otimização que visam à diminuição deste como objetivo da função de otimização (Župerl, Fornasaro *et al.* 2011).

RMSEP é calculado de acordo com a equação 2.12, onde $\hat{\boldsymbol{y}}$ representa a variável predita pelo modelo, $\boldsymbol{y}$ a variável medida e n o número de amostras contidas no grupo de amostras de teste. Este critério analisa o ajuste do modelo ao conjunto de dados de teste (ou calibração, no caso RMSEC), avaliando a reprodutibilidade dos dados.

$$\boldsymbol{RMSEP} = \sqrt{\sum_i^n (\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i)^2 / \boldsymbol{n}} \qquad (2.12)$$

O segundo critério para avaliação de modelos é o R^2, que indica o quanto modelos estatísticos se ajustam a dados reais. Este indicador fornece a informação do quanto os dados observados são replicados pelo modelo, quantificando a proporção da variação da resposta que é explicada pelas regressões em um modelo. No caso de modelos MLR, este índice pode ser classificado como sendo o quadrado do coeficiente de correlação entre as observações das variáveis e seus valores preditos.

O coeficiente de determinação assume valores entre 0 e 1, sendo R^2 igual a 1 indicativo de ajuste perfeito entre os dados observados e preditos, de forma que o modelo é capaz de explicar toda variabilidade da variável predita. De forma correlata, índices R^2 iguais a zero indicam a inexistência de correlação entre as variáveis medidas e preditas.

O cálculo do coeficiente de determinação é feito de acordo com a equação 2.13, onde $\bar{y}$ é a média do vetor de dados medidos ($\boldsymbol{y}$) (Everitt 2002).

$$\boldsymbol{R^2 = 1 - \sum_i^n (y_i - \bar{y})^2 \Big/ \sum_i^n (y_i - \hat{y}_i)^2} \qquad (2.13)$$

2.3 Métodos Espectroscópicos

Medidas baseadas na luz e outras formas de radiação eletromagnética são amplamente empregadas em caracterização de soluções e processos. As interações da radiação com a matéria são o objetivo de estudo da espectroscopia. Métodos espectroscópicos baseiam sua análise na medida da quantidade de radiação produzida ou absorvida por moléculas ou espécies atômicas de interesse. Estes métodos podem ser classificados de acordo com a região do espectro eletromagnético envolvido na medida (Settle 1997).

Os tipos de interação de maior interesse em espectroscopia envolvem transições entre diferentes níveis energéticos das espécies químicas. Outros tipos de interações, como reflexão, refração, espalhamento elástico, interferência e difração, são frequentemente relacionados às alterações das propriedades globais dos materiais do que com os níveis energéticos de moléculas ou átomos. Os tipos de interações que observamos dependem fortemente da energia da radiação empregada e o modo de detecção.

A Figura 2.3 apresenta as regiões do espectro eletromagnético que são empregadas em análises espectroscópicas. Nesta Figura também são expostos os tipos de transições atômicas e moleculares que resultam das interações da radiação com a amostra. Usualmente, métodos espectroquímicos que utilizam, não somente a radiação visível, como também a radiação ultravioleta e a infravermelha, são classificados como métodos ópticos, mesmo não empregando radiações na região do visível (no caso dos dois últimos tipos de radiação) (Skoog, Holler *et al.* 2007).

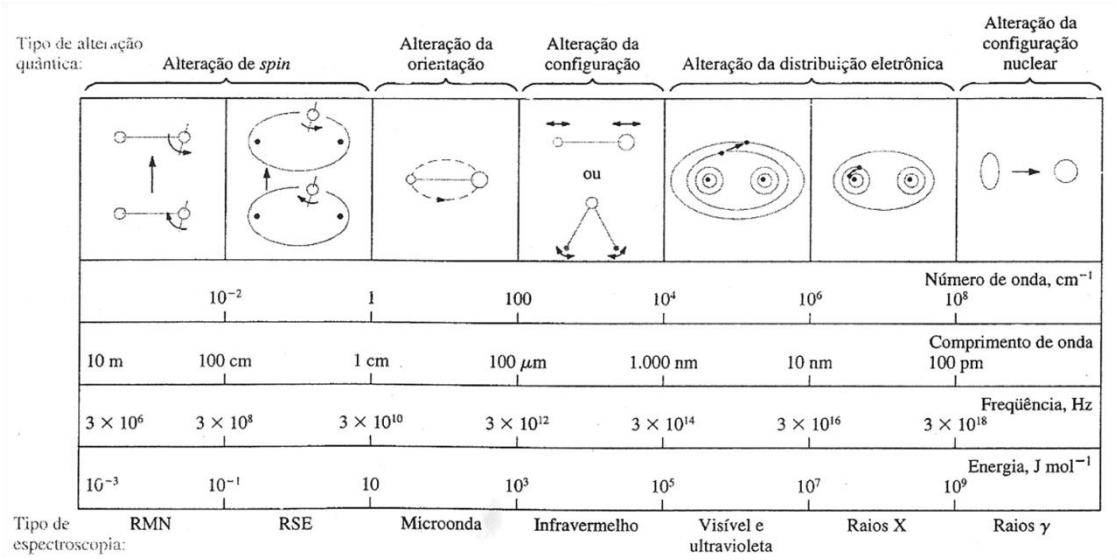

Figura 2.3: Regiões do espectro eletromagnético, salientando as possíveis alterações sofridas pelos analitos. **Fonte: Banwell (1983).**

Em análises espectroscópicas, a amostra é estimulada aplicando-se energia na forma de calor, energia elétrica, luz, partículas ou por uma reação química. Previamente ao estímulo, o analito se encontra predominantemente em seu estado fundamental. O estímulo então faz com que algumas espécies do analito sofram uma transição para um estado excitado de energia. Informações acerca do analito são obtidas através da análise da radiação emitida quando se retorna ao estado fundamental ou a quantidade de radiação eletromagnética absorvida decorrente da excitação.

A Figura 2.4 ilustra o processo envolvido na espectroscopia de emissão e de quimiluminescência. Neste exemplo, o analito é estimulado por calor, energia elétrica ou por uma reação química. A espectroscopia de emissão envolve métodos nos quais o estímulo é o calor ou energia elétrica, enquanto a espectroscopia de quimiluminescência refere-se à excitação do analito por meio de reação química.

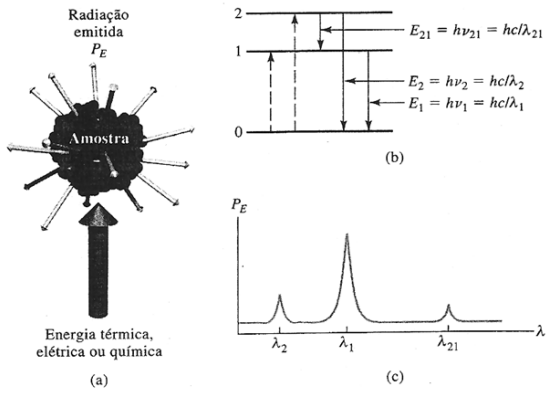

Figura 2.4: Processo de emissão ou quimiluminescência. (a) amostra é excitada pela aplicação de energia. (b) diagrama de níveis energéticos, onde as linhas pontilhadas representam a excitação não radiativa, enquanto as linhas cheias representam a perda de energia pela emissão de fótons. (c) espectro resultante mostrado como medida de potência radiante emitida em função do comprimento de onda. **Fonte: Skoog, *et al.* (2007).**

A Figura 2.5 ilustra o fenômeno responsável pela espectroscopia de fotoluminescência, onde a emissão de fótons é medida após a absorção da energia incidente. As formas mais importantes de fotoluminescência para os propósitos analíticos são as espectroscopias de fluorescência e fosforescência.

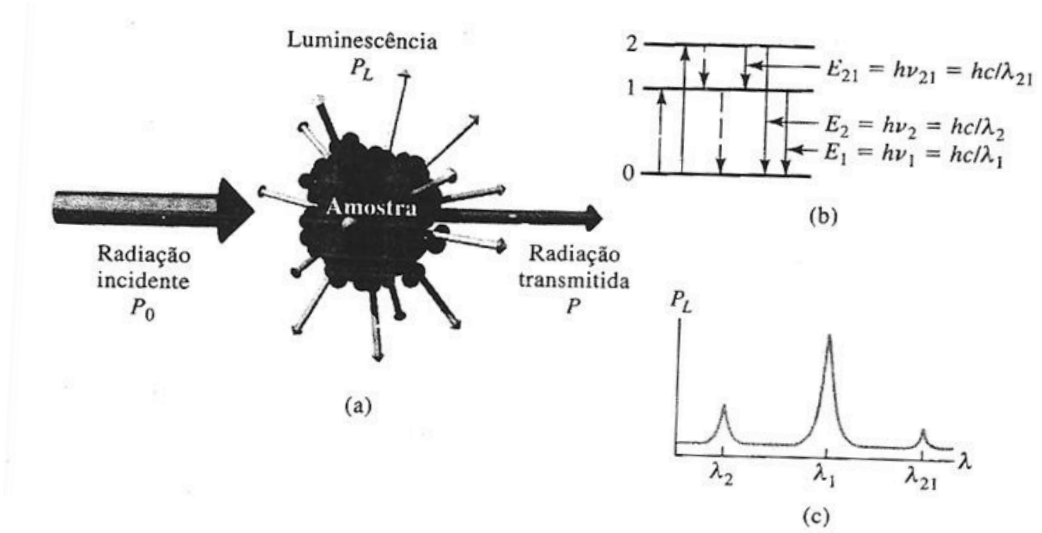

Figura 2.5: Fotoluminescência (fluorescência e fosforescência). (a) Absorção de radiação eletromagnética e dissipação de energia por emissão de radiação. (b) variação de energia dos níveis eletrônicos na etapa de absorção e emissão de fótons. (c) espectro correspondente aos comprimentos de onda emitidos associados às diferenças de energia entre os níveis. **Fonte: Adaptado de Skoog, *et al.* (2007).**

2.3.1 Espectroscopia de Fluorescência 2D

Durante os últimos 50 anos, houve um aumento significativo no uso da técnica de fluorescência em ciências biológicas, tornando a Espectroscopia de Fluorescência e a Fluorescência em tempo real ferramentas consideradas importantes em pesquisas bioquímicas e biofísicas (Lindemann, Marose *et al.* 1998).

Fluorescência é atualmente uma técnica dominante, aplicada intensivamente em biotecnologia, citometria de fluxo, diagnósticos médicos, sequenciamento de DNA, análise forense, análises genéticas, análises ambientais, dentre muitas outras aplicações (Marose, Lindemann *et al.* 1998, Hougaard, Lawaetz *et al.* 2013, Ammari, Bendoula *et al.* 2014, Bag, Yap *et al.* 2014, Hafuka, Yoshikawa *et al.* 2014, He, Dou *et al.* 2014, Omrani, Dudelzak *et al.* 2014, Valencia, Marín *et al.* 2014).

A principal característica desta técnica, que viabiliza sua aplicação em tantos processos, está em sua alta sensibilidade a detecção de metabólitos e componentes em solução, além da praticidade de operação, já que não é necessário o emprego de metodologias caras, ou o uso de traçadores radioativos para a maioria dos compostos bioquímicos medidos (Lakowicz 2006).

Luminescência é a emissão de luz a partir de qualquer substância e ocorre devido à estados eletrônicos excitados. Este fenômeno é formalmente dividido em três categorias

– fluorescência, fosforescência e quimiluminescência – dependendo da natureza dos estados excitados que dão origem a cada fenômeno (Ferd 1981).

A fluorescência e a fosforescência são fenômenos semelhantes, uma vez que para ambas a excitação é feita pela absorção de fótons. Como consequência, os dois fenômenos são frequentemente denominados pelo termo mais geral fotoluminescência. A fluorescência difere da fosforescência, pois as transições eletrônicas responsáveis pela fluorescência não envolvem mudança do spin eletrônico. Por outro lado, uma mudança de spin eletrônico acompanha as emissões por fosforescência, ocasionando tempos de vida maiores dos estados excitados. Usualmente, a fotoluminescência, seja como fluorescência ou fosforescência, ocorre em comprimentos de onda maiores que os da radiação de excitação (Skoog, Holler *et al.* 2007).

Em estados excitados do tipo singletos, o elétron presente no orbital excitado é pareado (pelo spin oposto) com o segundo elétron no estado orbital fundamental. Consequentemente, ao retornar ao estado fundamental, assumindo o valor determinado de spin, ocorre a rápida emissão de energia, na forma de fóton. A taxa de emissão de fluorescência é tipicamente 108 Hz, indicando que o tempo de vida típico da fluorescência é aproximadamente 10^{-9}s. É interessante avaliar o período de vida de 1 ns no contexto da velocidade da luz. A luz viaja cerca de 30 cm em um nano segundo. Muitos fluoróforos apresentam tempo de vida inferiores à nano segundos. Devido ao baixo tempo de emissão da fluorescência, tais medidas requerem equipamentos óticos e eletrônicos sofisticados, devido a grande complexidade em mensurar este tipo de medida (Wang, Yan *et al.* 2011).

O segundo tipo de luminescência existente, a fosforescência, é a emissão de luz de estados excitados do tipo tripletos, nos quais, o elétron presente no estado orbital excitado possui a mesma orientação spin que o elétron no estado orbital fundamental. Transmissões para o estado fundamental são proibidas e as taxas de emissão são lentas (de 1000 à 1 Hz). Devido a isso, fosforescência possui tempo de vida maior, da ordem de milissegundos a segundos. Após a exposição à luz, substâncias fosforescentes “brilham” por diversos minutos, enquanto os elétrons excitados retornam lentamente ao estado fundamental. Fosforescência não é usualmente observada em soluções à temperatura ambiente, isto porque nesta temperatura, existem diversos processos de desativação (relaxação vibracional, conversão interna e externa, etc) que competem com a emissão, como decaimentos não radioativos.

O processo que ocorre entre a absorção e a emissão de luz é usualmente ilustrado através dos Diagramas de Jablonski (Lakowicz 2006). Diagramas de Jablonski são geralmente utilizados como ponto de partida para discussões acerca do processo de absorção e emissão de luz, podendo ser utilizado de diversas formas distintas, para ilustrar processos moleculares que podem ocorrer em estados excitados (Danuta 1988).

A Figura 2.6 apresenta um diagrama parcial de níveis de energia (Diagrama de Jablonski) para uma molécula fotoluminescente típica. A linha horizontal mais grossa, na parte inferior do diagrama, representa a energia do estado fundamental da molécula, que é normalmente um estado singleto, e é denominada S_0. À temperatura ambiente, este estado representa as energias da maioria das moléculas em solução.

As linhas grossas na parte superior são os níveis de energia para os estados fundamentais vibracionais de três estados eletrônicos excitados. As duas linhas à

esquerda representam o primeiro (S_1) e o segundo (S_2) estados eletrônicos singleto. A linha à direita (T_1) representa a energia do primeiro estado eletrônico tripleto. Como normalmente ocorre, a energia do primeiro estado excitado tripleto é mais baixa que a energia do correspondente estado singleto.

Diversos níveis de energia vibracional estão associados a cada um dos quatro estados eletrônicos, conforme sugerido pelas linhas horizontais mais finas. Como exemplificado na Figura 2.6, as transições de absorção podem ocorrer do estado eletrônico fundamental singleto (S_0) para vários níveis vibracionais dos estados eletrônicos excitados singleto (S_1 e S_2). Vale salientar que a excitação direta do estado fundamental singleto para o estado excitado tripleto não é mostrada, uma vez que esta transição envolve uma mudança na multiplicidade, há uma probabilidade muito pequena da sua ocorrência. Uma transição de baixa probabilidade desse tipo é chamada transição proibida (Skoog, Holler *et al.* 2007).

O retorno ao estado fundamental usualmente ocorre para um estado vibracional fundamental superior ao S_0, o que rapidamente atinge o equilíbrio térmico. Este retorno a um estado excitado superior no nível de energia fundamental é o que dá origem à estrutura vibracional do espectro de emissão. Uma consequência interessante da emissão para níveis vibracionais de energia superiores a do nível fundamental é que o espectro de emissão é tipicamente um espelhamento do espectro de absorção da transição S_0 para S_1. Esta similaridade ocorre porque a excitação eletrônica não afeta de forma significativa a geometria nuclear, desta forma, o espaçamento entre os níveis de energia vibracional dos estados excitados são similares aqueles dos estados fundamentais. Como resultado, o espectro de emissão é similar ao espectro de absorção (Skoog, Holler *et al.* 2007).

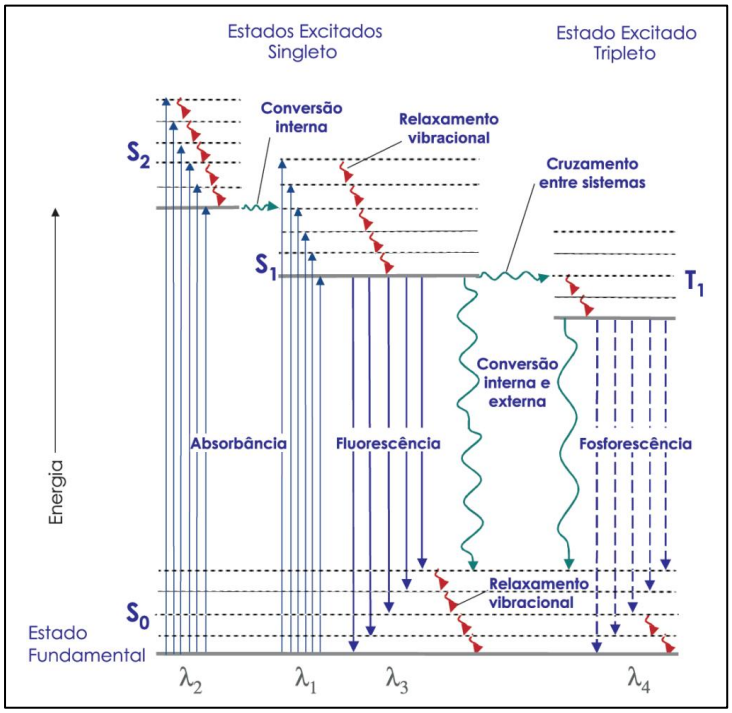

Figura 2.6: Diagrama parcial de níveis de energia para um sistema fotoluminescente. **Fonte: Sotomayor *et al.* (2008)**.

A característica mais atraente dos métodos luminescentes é sua inerente sensibilidade, com limites de detecção até três ordens de magnitudes melhores que aqueles encontrados na espectrometria de absorção. Existem casos específicos, sob condições controladas, onde de fato é possível detectar uma única molécula por Espectroscopia de Fluorescência. Outra característica muito interessante dos métodos fotoluminescentes é sua alta faixa linear de concentrações, que também é significativamente maior que a dos métodos de absorção (Whitford and Julien 2007).

O fenômeno da fluorescência dificilmente será resultado da absorção de radiação ultravioleta de comprimentos de onda inferiores à 250 nm, pois esse tipo de radiação é suficientemente energética para causar desativação dos estados excitados pela pré-dissociação ou dissociação. A maioria das moléculas orgânicas possuem ligações que podem ser rompidas por radiações com estas magnitudes de energia associadas (Tipler and Llewellyn 2008).

Fluorescência ocorre tipicamente em moléculas contendo grupos aromáticos funcionais com transições de baixa energia. Além destes, compostos contendo estruturas alifáticas e carbonilas alicíclicas ou estruturas de ligações duplas altamente conjugadas também podem apresentar fluorescência. A maioria dos hidrocarbonetos aromáticos não-substituídos (aromáticos substituídos são anéis aromáticos com estruturas anexas) apresentam fluorescência em solução e a eficiência quântica geralmente aumenta com o número de anéis e seu grau de condensação (Croney, Jameson *et al.* 2001).

Compostos heterocíclicos simples, como piridina, furano, tiofeno e pirrol, não apresentam fluorescência, entretanto, estruturas com anéis condensados geralmente fluorescem. Acredita-se que em compostos heterocíclicos nitrogenados a transição eletrônica de energia mais baixa envolva o sistema singleto sendo convertido rapidamente em tripleto, evitando a fluorescência, entretanto, a condensação dos anéis benzênicos a núcleos heterocíclicos resulta em um aumento da absorvidade da banda de absorção. Sendo assim, o tempo de vida de um estado excitado é mais curto nestas estruturas, observando-se a fluorescência para compostos como a quinolina, isoquinolina e indol (Skoog, Holler *et al.* 2007).

Os dados espectrais de fluorescência são geralmente apresentados como um espectro de emissão. O espectro de emissão de fluorescência é uma relação entre a intensidade de fluorescência versus comprimento de onda (nanômetros) ou número de onda (cm^{-1}).

A Figura 2.7 apresenta dois exemplos clássicos de espectros de emissão de fluorescência, sendo o primeiro associado a uma amostra de perileno e o segundo à quinina. Os espectros de emissão apresentam grandes variações, sendo dependentes da estrutura dos fluoróforos, bem como do solvente em que estes encontram-se dissolvidos. O espectro de alguns componentes, como por exemplo, perileno, apresentam significativa relação com os níveis vibracionais individuais dos estados de energia fundamentais e excitados, enquanto outros compostos, como quinina, apresentam espectro independente da estrutura vibracional (Lakowicz 2006).

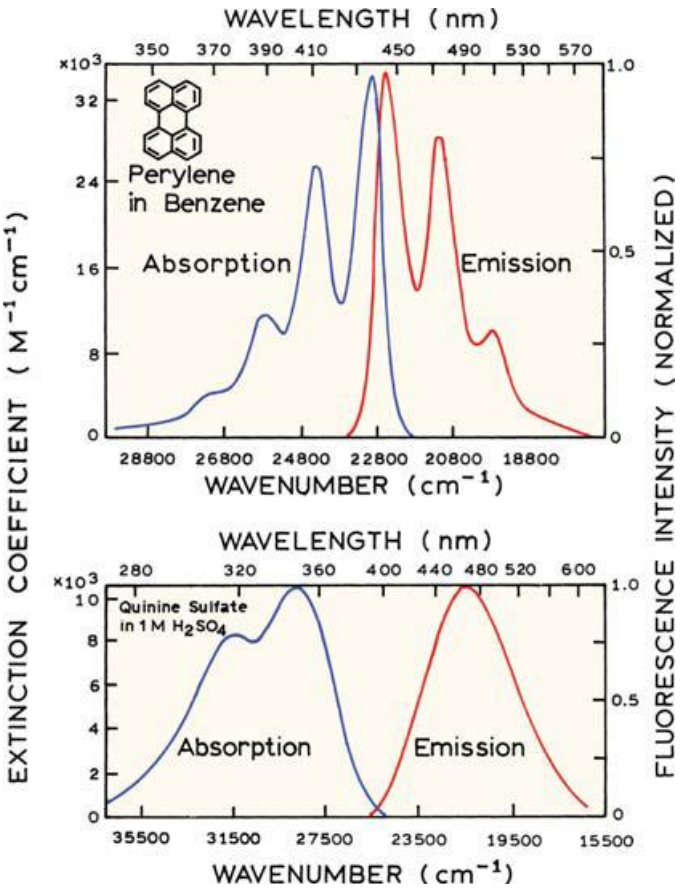

Figura 2.7: Espectros de absorção e emissão de fluorescência de perileno e quinina. **Fonte: Lakowicz (2006).**

O espectro de absorção, ou excitação, é obtido pela medida da intensidade de luminescência em um comprimento de onda fixo, enquanto o comprimento de onda de excitação varia. Como a primeira etapa para gerar fluorescência é a absorção de energia radiante para a geração de estados excitados, um espectro de absorção é essencialmente idêntico a um espectro de excitação, obtido para as mesmas condições. Os espectros de fluorescência e fosforescência, por sua vez, envolvem a excitação em um comprimento de onda fixo enquanto é feito o registro da intensidade de emissão em função do comprimento de onda (Mulchandani and Bassi 1995).

Os dados de Espectroscopia de Fluorescência também podem ser apresentados no que é chamado de Espectro de Luminescência Total ou Espectro Fluorescente 2D, apresentado na forma tridimensional ou em gráficos de contorno. Ambos gráficos mostram o sinal de luminescência em função dos comprimentos de onda de excitação e dos comprimentos de onda de emissão, de forma simultânea. O conjunto de dados que dá origem a este gráfico é usualmente chamado de Matriz de Excitação/Emissão. A Figura 2.8 apresenta um exemplo de cada um destes Espectros de Luminescência Total, onde em (a) é apresentado o espectro de fluorescência total de uma mistura de antraceno e de ovaleno, na forma de um gráfico em três dimensões, e em (b) é apresentado o gráfico de fluorescência total de 8-hidroxibenzopireno, na forma de curvas de contorno (Skoog, Holler *et al.* 2007).

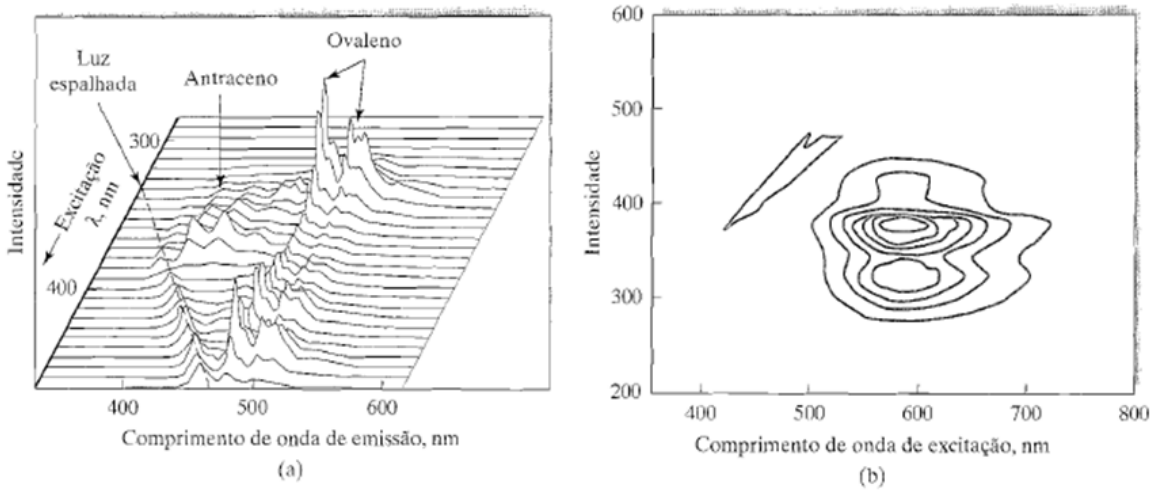

Figura 2.8: Espectros de Luminescência Total para uma mistura de antraceno e ovaleno (a), e para 8-hidroxibenzopireno (b). **Fonte: Skoog, *et al.* (2007)**.

Alguns instrumentos de luminescência permitem varrer simultaneamente os comprimentos de onda de excitação e de emissão com uma pequena diferença de comprimentos de onda entre eles. O espectro resultante é conhecido como Espectro Síncrono. Um sinal de luminescência é obtido apenas em comprimentos de onda onde a excitação e a emissão ocorrem para a diferença de comprimentos de onda escolhida. O Espectro Síncrono também pode ser obtido através do Espectro de Luminescência Total, selecionando dados específicos deste (Skoog, Holler *et al.* 2007).

Os componentes presentes em equipamentos destinados a medir fotoluminescência são similares àqueles encontrados em fotômetros ou espectrofotômetros ultravioleta-visível. A Figura 2.9 apresenta um esquema genérico do arranjo destes componentes em um fluorômetro ou espectrofluorômetro. De forma genérica, quase a totalidade de equipamentos deste tipo empregam a ótica de duplo feixe em suas medidas, onde o feixe de luz superior passa inicialmente por um seletor de excitação de comprimento de onda (filtro ou monocromador), o qual transmite a radiação responsável pela excitação de fluorescência, mas exclui ou limita a radiação de comprimento de onda de emissão fluorescente.

A fluorescência é emitida pela amostra em todas as direções, sendo mais convenientemente observada em ângulos retos ao feixe de excitação, minimizando as contribuições por espalhamento e da radiação intensa da fonte. A radiação emitida, equivalentemente à excitada, passa por um seletor de comprimento de onda (filtro ou monocromador) que isola a emissão da fluorescência. Esta radiação isolada atinge um foto transdutor que converte a intensidade da luz em um sinal elétrico utilizado na medida.

O segundo feixe oriundo da fonte de energia radiante, gerado após o sinal oriundo do seletor de excitação passar através de um divisor de feixes, usualmente um espelho semitransparente, é focado através de um atenuador que reduz sua potência para um valor próximo daquele da radiação de fluorescência, reduzindo cerca de cem vezes sua intensidade. Este feixe atenuado atinge um segundo transdutor, sendo convertido em um sinal elétrico. Os componentes eletrônicos, associados em conjunto com um sistema de análise de dados, processam os dois sinais de forma a calcular a razão da intensidade de emissão de fluorescência para a intensidade da fonte de excitação, desta forma é possível

cancelar os efeitos oriundos de flutuações na intensidade da fonte, e produzir o espectro de fluorescência, ou os dados para um único comprimento de onda (Skoog, Holler *et al.* 2007).

Fluorômetros são equipamentos que empregam apenas filtros para fazer a seleção dos comprimentos de onda de excitação e emissão, enquanto que Espectrofluorômetros empregam dois monocromadores para isolar os comprimentos de onda (Omary and Patterson 1999).

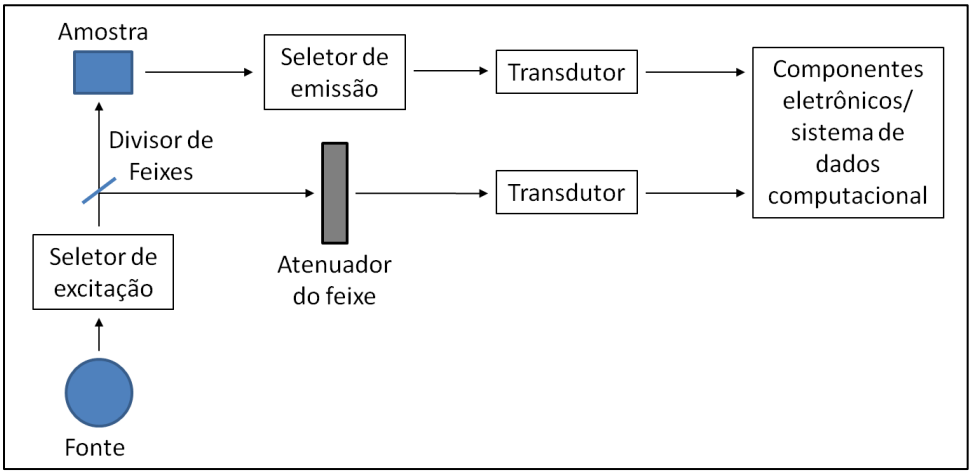

Figura 2.9: Esquema da estrutura genérica dos componentes de um equipamento Fluorômetro ou Espectrofluorômetro. **Fonte: Skoog, *et al.*, 2007**.

Os fluorômetros de filtro são uma forma relativamente barata para o desenvolvimento de análises quantitativas de fluorescência, sendo os filtros de absorção ou interferência utilizados para limitar os comprimentos de onda de excitação e radiação emitida. Estes equipamentos são considerados em sua grande maioria compactos, robustos e de fácil utilização (Roberts, Dancik *et al.* 2011).

Os espectrofluorômetros possuem a característica de permitir a obtenção de espectros de excitação de fluorescência ou ainda espectros de emissão de fluorescência, já que sua configuração estrutural permite mensurar espectros de excitação onde a intensidade de emissão de fluorescência pode ser mantida constante em um único comprimento de onda enquanto varre os comprimentos de onda de excitação. Além de que, fazendo as devidas correções para as variações de intensidade na saída da fonte e a resposta do detector como uma função do comprimento de onda, podem ser obtidos espectros de excitação absolutos, que são similares aos espectros de absorção (Skoog, Holler *et al.* 2007).

Diversos fabricantes de espectrofluorômetros disponibilizam equipamentos com a capacidade de obtenção dos dois espectros, excitação e emissão. O projeto estrutural destes é um pouco mais complexo que o projeto estrutural de um fluorômetro, já que a utilização de monocromadores, como monocromadores de rede, implica na utilização de um maior número de componentes óticos, como lentes e espelhos. Apesar desta diferença em termos de componentes, análogo ao fluorômetro, no espectrofluorômetro, a radiação oriunda da fonte de excitação é direcionada ao monocromador de excitação, sendo a radiação resultante deste componente dividida através da utilização de um divisor de feixes, onde um dos feixes é direcionado à fotomultiplicadora de referência, enquanto o segundo feixe é direcionado à amostra. A radiação fluorescente resultante,

após a dispersão pelo monocromador de emissão é detectada pela segunda fotomultiplicadora (Lakowicz 2006).

Espectrofluorômetros como o descrito anteriormente fornecem espectros adequados para análises quantitativas, entretanto, não serão obrigatoriamente comparáveis aos espectros obtidos por outros equipamentos, uma vez que a saída não depende apenas da intensidade de fluorescência, mas também de outras características da lâmpada, do transdutor e dos monocromadores. Todas estas características instrumentais variam com o comprimento de onda e diferem de instrumento para instrumento.

2.3.2 NIR – Espectroscopia no Infravermelho Próximo

O método de espectroscopia no infravermelho próximo (*NIRS – Near Infrared Spectroscopy* ou simplesmente NIR – *Near Infrared*) possui algumas vantagens que motivam sua utilização na caracterização de processos, dentre as quais podem ser citadas a velocidade de medida quando comparada a outros processos (usualmente medidas são tomadas em menos de 1 segundo), sendo uma técnica não destrutiva e que na grande maioria dos casos não necessita de preparação da amostra. Além destas vantagens, esta técnica é considerada altamente versátil. Caso as amostras possuam ligações C-N, N-H ou O-H, e caso a concentração do analito ultrapasse 1% da composição total, resultados promissores podem ser esperados quase que na totalidade dos casos (Alves and Poppi 2013).

A técnica de espectroscopia no infravermelho é classificada como uma técnica de espectroscopia vibracional, na qual também é incluída a técnica de espectroscopia Raman. Estes dois tipos de espectroscopia fornecem o mesmo tipo de informação molecular, sendo um método complementar ao outro.

No espectro eletromagnético, a região do infravermelho é dividida em três partes: (a) Infravermelho Próximo (NIR), classificado como região das harmônicas e compreendido entre 0,8 – 2,5 µm (12500 – 400 cm^{-1}), (b) Infravermelho Médio (MIR), classificado como região de vibração-rotação e compreendido entre 2,5 – 50 µm (400 – 200 cm^{-1}) e (c) Infravermelho Distante, classificado como região de rotação e compreendido entre 50 – 1000 µm (200 – 10 cm^{-1}).

Diversas são as vantagens de utilização da metodologia NIR e MIR para caracterização de processos. A Tabela 2.1 apresenta uma comparação entre as características qualitativas das técnicas de espectroscopia vibracional MIR e NIR. A partir desta tabela é possível escolher, baseado nas características almejadas no processo de caracterização, qual das técnicas apresenta melhor viabilidade para aplicação.

Tabela 2.1: Comparação das características qualitativas de MIR e NIR. **Fonte: Adaptado de Pasquini (2002).**

	MIR	NIR
	Vibrações Fundamentais	Sobretons e Combinações
Qualitativa	Excelente (estrutura)	Ruim (identidade)
Quantitativa	Excelente	Excelente

Intensidade	Alta	Baixa
Espessura da amostra	Muito pequena	Grande
Materiais	KBr/NaCl	Quartzo/Vidro
Sinal/Ruído	$<10^4$	$>>10^4$
Refletância	Satisfatória	Excelente

A espectroscopia NIR se baseia no fato de que as ligações químicas das substâncias possuem frequências de vibração específicas, que correspondem a níveis de energia das moléculas, também chamados de níveis vibracionais. Estes níveis são dependentes da forma da superfície de energia potencial da molécula, de sua geometria, da massa dos átomos que a compõe e do seu acoplamento vibracional.

Ao ser exposta à radiação eletromagnética com a mesma quantidade de energia de um de seus níveis vibracionais, esta energia será absorvida desde que certas condições sejam atendidas. Neste caso, para que esta absorção seja lida no espectro NIR, a molécula precisa sofrer uma variação em seu momento dipolar no decorrer do processo de absorção da energia.

Nesta técnica, os espectros são consequência das bandas harmônicas ou de combinação das frequências fundamentais. As transições envolvendo harmônicos são "proibidas", mas elas são observadas devido à anarmonicidade dos osciladores reais. As chamadas bandas proibidas são de 10 a 1000 vezes mais fracas do que as bandas fundamentais.

A espectroscopia NIR comporta a região de números de onda adjacentes à MIR e se estende até a região do visível. A absorção NIR é baseada em sobretons (*overtone*) e combinação das vibrações da molécula investigada e suas baixas probabilidades de transição. Uma vez que as intensidades de absorção decrescem sucessivamente na direção do MIR ao visível, existe a possibilidade de ajuste na espessura da amostra, passando da ordem de grandeza de milímetros até centímetros, dependendo da classificação do sobretom. Esta é uma diferença significativa em comparação a espectroscopia MIR e Raman, onde a intensidade dos sinais das vibrações fundamentais variam irregularmente no intervalo de frequência e dependem exclusivamente das condições de excitação vibracional molecular individual (Siesler, Ozaki *et al.* 2002).

Estruturalmente, os equipamentos de espectroscopia NIR são similares aos instrumentos para medidas de espectroscopia UV-visível e MIR. Os componentes base são a fonte luminosa, detector e o elemento dispersivo, que pode ser um prisma ou então uma rede de difração. Equipamentos que trabalham com infravermelho por transformada de Fourier utilizam de interferômetros, especialmente para comprimentos de onda superiores à 1000 nm.

Em equipamentos de espectroscopia NIR, usualmente, a fonte luminosa é uma lâmpada alógena de banda larga com janela de quartzo e aproveitamento de radiação entre 0,8 e 2,5 µm. A dispersão da radiação é feita com redes holográficas cortadas com laser e movidas por motores de movimento descontínuo (motor de passo). Diodos emissores de luz também estão sendo empregados para esta função. Além de

apresentarem um tempo de vida útil elevada, possuem grande estabilidade espectral, além de reduzido consumo de energia.

No caso da espectroscopia de infravermelho por transformada de Fourier, a técnica visa a coleta dos espectros de forma mais rápida, evitando a variação da frequência de radiação infravermelha na qual a amostra é exposta na técnica convencional. Nesta técnica, a amostra é irradiada com todos os comprimentos de onda da radiação de infravermelho da faixa desejada e o sinal resultante, denominado interferograma, é então decomposto através da transformada de Fourier, resultando em um espectro equivalente ao de espectroscopia dispersiva convencional.

Os equipamentos que utilizam da técnica de espectroscopia de infravermelho por transformada de Fourier são mais acessíveis economicamente devido a facilidade de obtenção de interferômetros, frente à monocromadores. A vantagem do ganho de tempo deste equipamento se deve a coleta simultânea de todas as frequências, o que permite a coleta de diversas medidas da mesma amostra e a avaliação da media destas, ao invés de medidas singulares, aumentando a sensibilidade da análise.

Dois tipos de detectores são utilizados para cobrir toda a faixa útil do espectro: (a) detectores de silício para a região entre 0,8 e 1,1 μm, e de (b) sulfeto de chumbo para a região compreendida entre 1,1 e 2,5 μm.

Assim como todo método analítico de caracterização, a técnica de espectroscopia NIR também está sujeita à presença de ruídos nas medidas. De forma genérica, toda medida analítica é composta pelo sinal, que contém a informação qualitativa e quantitativa desejada, e o ruído, que nada mais é que uma fonte de informação estranha que afeta o limite de detecção, precisão e exatidão do método. Em NIR, a fonte de ruído pode ser oriunda de duas vertentes, o ruído químico e o ruído instrumental.

O ruído químico resulta de uma grande quantidade de variáveis que afetam os sistemas químicos, principalmente variações físico-químicas não detectadas como temperatura, pressão, pH, umidade, etc. O ruído instrumental por sua vez, é associado aos componentes do equipamento, como fontes, transdutores, processadores de sinal, etc. Desta forma, para elevar o nível das medidas em sistemas de espectroscopia NIR, assim como nas demais técnicas analíticas, é aconselhada a utilização de elementos de hardware como filtros, moduladores e detectores sincronizados associados a sistemas de software capazes de permitir a extração de sinais ambientais que impõe ruídos as medidas, principalmente filtros digitais.

A espectroscopia NIR é sensível de forma primária aos grupos funcionais das moléculas do analito e em sequência a efeitos decorrentes dos níveis atômicos, microscópicos e macroscópicos das amostras. Com relação aos fatores químicos que afetam os espectros vibracionais, devem ser mencionados: a posição da banda de energia, que diz respeito a constante de forca da ligação e massa dos átomos presentes nas moléculas, à intensidade da banda de energia, fazendo relação com a alteração do momento dipolo que acompanha as vibrações e por fim os fatores referentes à simetria das moléculas, uma vez que a simetria determina se a banda é classificada como ativa ou inativa, além de influenciar na probabilidade de se observar combinações e ressonância entre os modos vibracionais presentes. De forma genérica, a espectroscopia NIR é sensível a qualquer fator que afete a massa atômica, forca de ligação, momento dipolo ou simetria da molécula.

Efeitos primários são os mais significantes para os espectros em questão, sendo aqueles correlacionados aos grupos funcionais e diretamente associados a massa atômica, forças de ligação e anarmonicidade. Um exemplo deste efeito é o fato das bandas do primeiro sobretom de estiramento da ligação C-H ocorrer em frequências mais altas do que para a ligação N-H, devido ao fato da massa do C ser menor que a do N.

Os efeitos secundários associados aos níveis atômicos são significativos quando a vizinhança do grupo funcional apresenta características de forte doador ou receptor de elétrons. Estas características afetam a força da ligação e o momento de dipolo do grupo de interesse. Por sua vez, os efeitos secundários de nível microscópico mais comuns que afetam os espectros NIR estão relacionados à cristalinidade dos materiais, ou seja, são referentes a forma como as moléculas se ordenam espacialmente em uma grande extensão de material. Por fim, os fatores secundários de nível macroscópicos são aqueles associados a variações de temperatura e submissão das amostras a efeitos mecânicos (ex. aplicação de tensão e alteração da distribuição dos estados conformacionais das moléculas).

Capítulo 3 – Estudos de Caso e PSCM

Neste capítulo é apresentada a descrição dos dados utilizados como estudo de caso para as análises das metodologias quimiométricas propostas neste trabalho. Além destas, também é apresentada a descrição das implementações relativas ao método de otimização global Colônia de Formigas (ACO), utilizada como base dos métodos propostos.

Dois conjuntos de dados distintos são utilizados no decorrer deste trabalho. O primeiro deles, composto por dados de fermentação de glicose com a levedura *Saccharomyces cerevisiae* visa avaliar o desempenho dos métodos apresentados no estudo e caracterização de dados de processos reais. Os dados em questão são relativos a duas fermentações da referida levedura no processo de produção de etanol e posterior consumo deste até seu consumo total. O acompanhamento destas fermentações é feito de forma *off-line*, através da coleta de amostras e posterior análise laboratorial, fornecendo informações relativas às variáveis de estado de interesse do sistema, isto é, concentrações de substrato (glicose), biomassa e produto (etanol). Durante o decorrer das fermentações, espectros de fluorescência 2D foram coletados, sendo estes dados utilizados no estudo deste trabalho. O segundo conjunto de dados utilizados neste trabalho visa à avaliação das técnicas apresentadas para a caracterização de amostras aleatórias de sistemas específicos. Para tanto, este conjunto de dados é formado por amostras de farinha de trigo, de marcas distintas, coletadas em estabelecimentos comerciais aleatórios, caracterizadas analiticamente através de análises farinográficas e espectralmente através da técnica de espectroscopia de infravermelho próximo (NIR).

3.1 Fermentação de Glicose utilizando *S. cerevisiae*

O conjunto de dados relativo a processos reais, utilizado neste trabalho, é constituído por dados experimentais fornecidos pelo professor B. Hitzmann, enquanto este se encontrava no *Institut für Technische Chemie, University of Hannover,* Alemanha. Nos experimentos em questão, foi utilizada a levedura *S. cerevisiae* da linhagem H620, crescendo em biorreator com volume de 1,5 L à temperatura constante de 30 $^{\circ}$C e pH 5,5 com a utilização de suplemento de meio Schatzmann e glicose.

Durante os cultivos, espectros de fluorescência foram coletados a cada 6 minutos pelo equipamento BioView-espectrômetro (Delta Light & Optics, Denmark) como descrito por Stärk *et al.* (2002). Cada espectro é constituído de 150 pares de comprimento de onda de excitação/emissão, composto como segue: 15 filtros na região de 270nm à 550nm para excitação e 15 filtros na região de 310nm à 590nm para emissão com largura de banda de 20nm cada e coletado equidistante de 20nm.

Paralelamente à coleta de dados *on-line* por espectroscopia de fluorescência, os cultivos foram caracterizados de forma *off-line* por Cromatografia Líquida de Alta Eficiência (HPLC), de forma a possibilitar o acompanhamento da evolução das concentrações de glicose e etanol, e por análise gravimétrica para a concentração de biomassa. As Tabelas 3.1 e 3.2 apresentam os conjuntos de dados *off-line* obtidos para o ensaio fermentativo 1 e ensaio fermentativo 2, respectivamente.

Tabela 3.1: Conjunto de dados experimentais *off-line* para o ensaio fermentativo 1. **Fonte: Solle *et al.* (2003).**

Tempo [h]	Biomassa [g/L]	Etanol [g/L]	Glicose [g/L]
0,0	0,3	0,40	30,00
2,5	0,5	0,60	28,80
5,0	1,3	2,10	25,20
6,5	2,3	4,30	17,90
8,0	4,3	9, 40	3,54
9,5	5,1	9,40	0,01
14,1	6,2	6,30	0,00
16,0	7,3	4,10	0,00
18,1	8,9	1, 50	0,00
19,4	9,8	0,01	0,00

Tabela 3.2: Conjunto de dados experimentais *off-line* para o ensaio fermentativo 2. **Fonte: Solle *et al.* (2003).**

Tempo [h]	Biomassa [g/L]	Etanol [g/L]	Glicose [g/L]
0,0	0,2	0,40	30,0
2,5	0,6	0, 80	28,0
5,0	1,3	2,30	21,6
6,5	2, 5	5,00	14,4
8,3	4,7	9,90	0,0
9,5	4,9	9,30	0,0
14,1	6,1	5,80	0,0
16,0	7,3	3,30	0,0
18,1	8,9	0,70	0,0
19,0	9,2	0,01	0,0

Devido à baixa quantidade de informações *off-line* disponíveis, frente à grande quantidade de espectros de fluorescência obtidos para cada fermentação (cerca de 190 espectros para cada ensaio), foi feito uso de ferramentas de simulação dinâmica de processos para obtenção de dados simulados de processo *off-line* correspondentes aos instantes de coleta dos dados de Espectroscopia de Fluorescência.

Inicialmente, foram gerados dados de caracterização off-line, obtidos por interpolação linear, para os três estados avaliados nos dois ensaios fermentativos, com amostragem equivalente àquela utilizada na coleta de dados de Espectroscopia de Fluorescência 2D (Figura 3.1). Este conjunto de dados foi o utilizado para o ajuste e teste de parâmetros do modelo dinâmico para o referido processo, com o objetivo de obter dados *off-line* simulados e representativos dos ensaios.

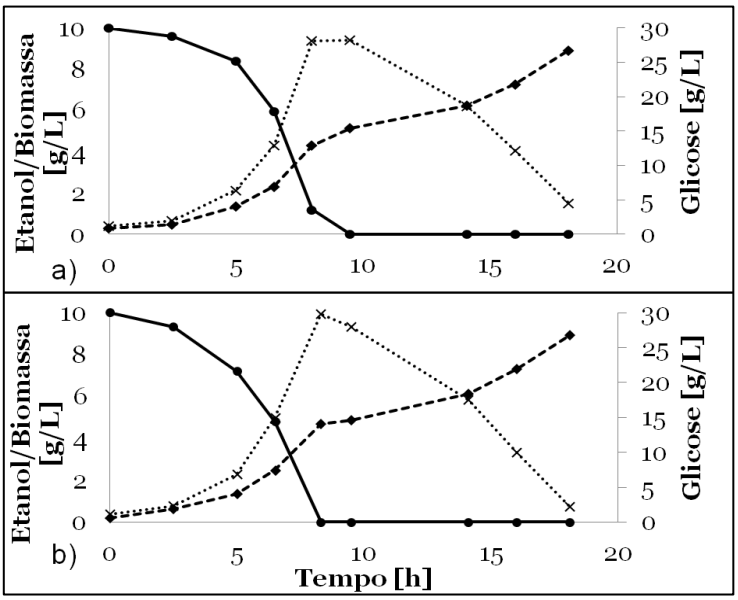

Figura 3.1: Interpolação de dados *off-line* dos estados fermentativos. (a) ensaio fermentativo 1 e (b) ensaio fermentativo 2. (×) Etanol, (●) Glicose e (♦) Biomassa.

Quando células de *S. cerevisiae* são expostas à glicose, em um meio propício para seu crescimento, produzem etanol e biomassa, mesmo em condições aeróbicas (efeito Crabtree) e pode ser observado o padrão de crescimento diáuxico da população (Zang, Scharer *et al*. 1997, Pratap R 2003, Pratap 2003). Este processo, quando operado em configuração batelada, pode ser descrito pelo sistema de equações diferenciais 3.1 à 3.3, conforme Zang *et al.* (1997).

$$\frac{dG}{dt} = -\mu_G \frac{X}{Y_{GX}} \tag{3.1}$$

$$\frac{dE}{dt} = \mu_G \frac{X}{Y_{GE}} - \mu_E \frac{X}{Y_{EX}} \tag{3.2}$$

$$\frac{dX}{dt} = \mu_E X + \mu_G X \tag{3.3}$$

G, *E* e *X* são respectivamente as concentrações de glicose, etanol e biomassa no meio fermentativo. Os termos μ_G e μ_E representam as taxas de crescimento específica em glicose e etanol respectivamente. Y_{GX}, Y_{GE} e Y_{EX} são os rendimentos específicos de glicose em biomassa, glicose em etanol e etanol em biomassa.

Matematicamente, o crescimento diáuxico pode ser considerado como uma descontinuidade no modelo dinâmico. Este fenômeno implica na biomassa utilizar, como fonte de carboidratos primários, a glicose presente no meio, e na falta desta, passar a utilizar o etanol produzido anteriormente como fonte de energia para a manutenção celular.

No modelo dinâmico, o crescimento diáuxico implica em μ_G ser maior que zero apenas quando existe glicose no meio, e consequentemente, μ_E é igual a zero, não havendo crescimento em etanol (repressão de crescimento por glicose). Quando a concentração de glicose é igual à zero, ou seja, toda glicose adicionada no início do processo é consumida e o etanol passa a ser a única fonte de carbono disponível, μ_E passa a ser maior que zero até que todo etanol seja consumido e μ_G passa a valer zero.

O fato dos valores de μ_G e μ_E apresentarem descontinuidade em função da concentração de glicose no meio fermentativo, faz com que tanto simulações deste modelo, bem como ajuste de parâmetros, tornem-se procedimentos complexos com maiores incertezas associadas. Para resolver este problema, foram propostas novas equações para as taxas de crescimento, de forma a possibilitar que o mesmo capture a mudança no metabolismo das leveduras e apresente maior representatividade do sistema fermentativo, retirando a descontinuidade apresentada pelo modelo dinâmico original e tornando-o contínuo.

A modificação estrutural proposta consiste em não considerar μ_G e μ_E como constantes, como proposto no trabalho de Solle *et al.* (2003) (neste trabalho os autores utilizam operadores lógicos para promover a simulação da mudança de metabolismo apresentada pelo microrganismo), mas sim, funções da concentração de glicose no meio ($\mu_G(G)$ *e* $\mu_E(G)$). As equações 3.4 e 3.5 apresentam as funções propostas para as taxas de crescimento, enquanto a Figura 3.2 simula o exemplo de comportamento esperado destas taxas de crescimento, em função da concentração de glicose do meio, mantendo os parâmetros *a*, *b*, μ_{Gm} e μ_{Em} iguais à 1. O objetivo desta Figura é exemplificar forma de transição das taxas de crescimento através das modificações modeladas.

$$\boldsymbol{\mu_G = \mu_{Gm}\left(0,5 - 0,5tanh\left(a(b - G)\right)\right)} \tag{3.4}$$

$$\boldsymbol{\mu_E = \mu_{Em}\left(0,5 - 0,5tanh\left(a(G - b)\right)\right)} \tag{3.5}$$

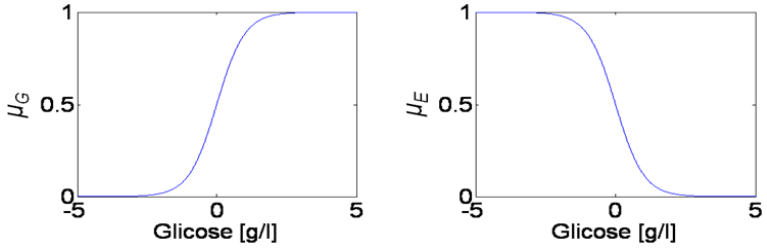

Figura 3.2: Comportamento de μ_G e μ_E em função da concentração de glicose, mantendo os parâmetros *a*, *b*, μ_{Gm} e μ_{Em} constantes e iguais à 1.

As etapas de simulação dinâmica das equações diferencias do modelo, para a obtenção da evolução das variáveis de estado no tempo, é feita através do método de integração numérica Runge-Kutta, implementado na função *fminsearch* do MatLab 5.3. Nestas simulações, os coeficientes de conversão são mantidos constantes em valores determinados experimentalmente para os experimentos em questão, e obtidos no trabalho de Solle *et al.* (2003) (Y_{GX} = 0,167 $g_{células}/g_{glicose}$, Y_{GE} = 0,5 $g_{etanol}/g_{glicose}$, Y_{EX} = 0,333 $g_{células}/g_{etanol}$).

A modificação proposta para as taxas de crescimento proporciona a captura da mudança de metabolismo pelas leveduras, sem a necessidade de utilização de operadores lógicos (estratégia utilizada por Solle *et al.* (2003)). Os parâmetros *a* e *b* indicam, respectivamente, o quão abrupta é a mudança de metabolismo, e para qual valor de concentração de glicose essa mudança deve ocorrer. As variáveis μ_{Gm} e μ_{Em} fazem referência direta à μ_G e μ_E presentes na estrutura inicial de modelo. O modelo dinâmico resultante das modificações propostas é apresentado nas equações 3.6 à 3.8.

$$\frac{dG}{dt} = -\frac{X\mu_{Gm}\left(0{,}5-0{,}5tanh\left(a(b-G)\right)\right)}{Y_{GX}} \tag{3.6}$$

$$\frac{dE}{dt} = \frac{X\mu_{Gm}\left(0{,}5-0{,}5tanh\left(a(b-G)\right)\right)}{Y_{GE}} - \frac{X\mu_{Em}\left(0{,}5-0{,}5tanh\left(a(G-b)\right)\right)}{Y_{EX}} \tag{3.7}$$

$$\frac{dX}{dt} = X\mu_{Em}\left(0{,}5 - 0{,}5tanh\left(a(G-b)\right)\right) + X\mu_{Gm}\left(0{,}5 - 0{,}5tanh\left(a(b-G)\right)\right) \tag{3.8}$$

Para a validação da estrutura de modelo proposta, dados interpolados do ensaio fermentativo 1 (Figura 3.1a) são utilizados para ajuste dos parâmetros do modelos, que são posteriormente utilizados para simulação do ensaio fermentativo 2 (Figura 3.1b). A Tabela 3.3 apresenta os valores dos parâmetros obtidos no ajuste do modelo, enquanto a Figura 3.3a apresenta o resultado do modelo ajustado aos dados do ensaio fermentativo 1 e a Figura 3.3b apresenta os resultados da simulação deste modelo com dados iniciais do ensaio fermentativo 2.

Tabela 3.3: Parâmetros do modelo dinâmico ajustados com dados interpolados do ensaio fermentativo 1.

Parâmetro	Valor	Unidade
μ_{Gm}	0,3792	1/h
μ_{Em}	0,0587	1/h
a	1,3968	L/g
b	2,2136	g/L

A Figura 3.3 comprova a eficiência do modelo proposto para fermentações de glicose com a levedura *S. cerevisiae*. A boa representatividade dos dados reais, apresentada pela simulação das variáveis de estado do ensaio fermentativo 2, validam a estrutura proposta, bem como as constantes ajustadas para o modelo dinâmico. É interessante ressaltar que o comportamento diáuxico, apresentado por este microorganismo, é capturado pelo modelo, e os ajustes obtidos para μ_{Gm} e μ_{Em} são equivalentes aos publicados por Solle *et al.* (2003), para o modelo de Zang *et al.*, (1997), comparados respectivamente à μ_G=0,41 L/h e μ_E=0,048 L/h.

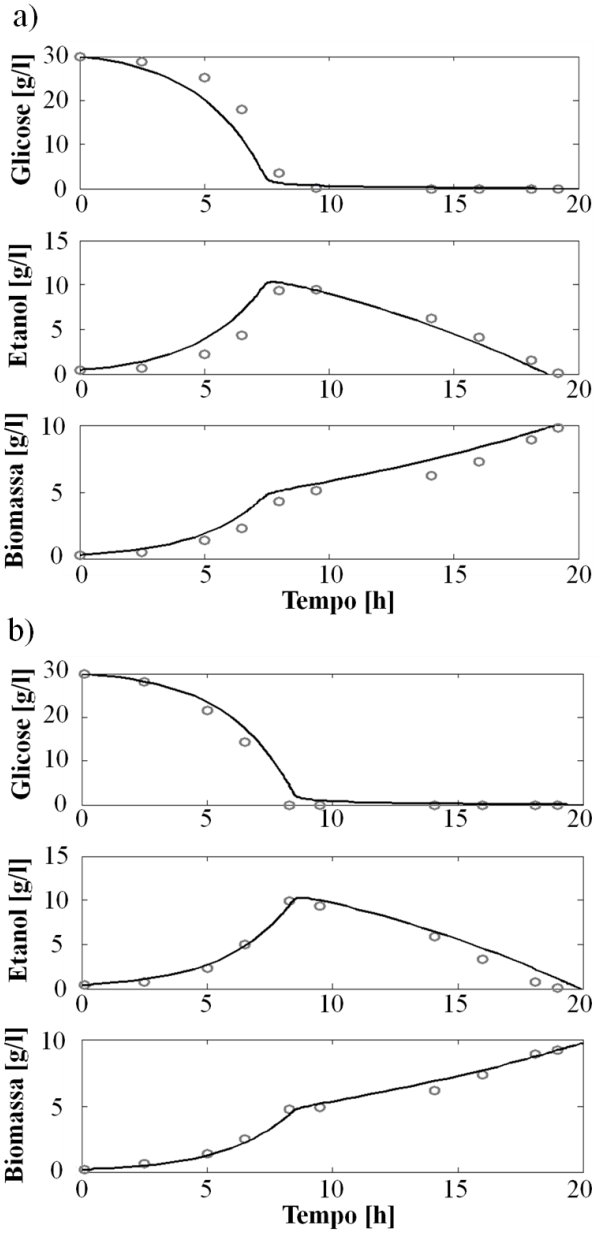

Figura 3.3: Resultados da validação da estrutura de modelo dinâmico proposta para fermentações batelada de glicose usando *S. cerevisiae.* (a) Etapa de calibração dos parâmetros do modelo usando dados do ensaio fermentativo 1 e (b) modelo ajustado simulando dados do ensaio fermentativo 2. (—) Modelo, (○) dados *off-line*.

Uma vez que a estrutura de modelo dinâmico proposto para o sistema em questão esta validada, é realizado o ajuste de parâmetros do modelo utilizando simultaneamente os dois conjuntos de dados de fermentação. Assim, o modelo ajustado resultante é correlacionado de forma equivalente a ambas as fermentações, sendo igualmente representativo para os dois conjuntos de dados experimentais.

O modelo resultante deste ajuste simultâneo possui os seguintes parâmetros otimizados: μ_{Gm} = 0,3403 l/h, μ_{Em} = 0,0597 l/h, a = 4,1011 l/g e b = 0,925 g/L, sendo o resultado da simulação das taxas de crescimento ajustadas, em função da concentração de glicose, apresentado na Figura 3.4, enquanto a Figura 3.5 apresenta a simulação

dinâmica das variáveis de estado deste processo fermentativo, frente os dados experimentais de cada batelada.

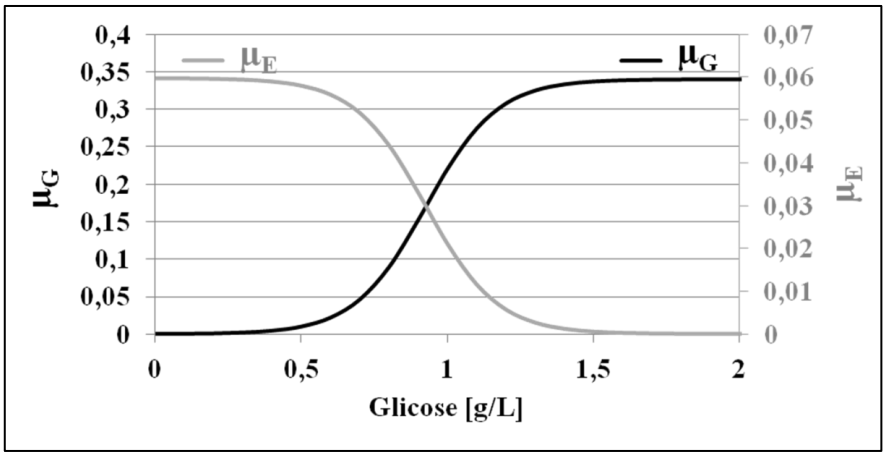

Figura 3.4: Simulação das taxas de crescimento μ_G e μ_G em função da concentração de glicose no meio reacional, usando parâmetros ajustados iguais a: μ_{Gm} = 0,3403 l/h, μ_{Em} = 0,0597 l/h, a = 4,1011 l/g e b = 0,925 g/L.

A Figura 3.4 mostra que o intervalo de conversão entre as taxas de crescimento, ou seja, a mudança ocorrida no metabolismo das leveduras que deixa de consumir glicose, para passar a consumir o etanol produzido, uma vez que o meio reacional encontra-se pobre em glicose, não ocorre de forma instantânea e não ocorre em concentração de glicose igual à 0 g/L, como modelado inicialmente por Zang *et al.*, (1997).

O ajuste de parâmetros obtido para o modelo modificado, mostra que esta transição entre as taxas de crescimento ocorre no intervalo de valores de concentração de glicose compreendidos, aproximadamente, entre 0,5 g/L e 1,5 g/L. Caso esta transição fosse realmente instantânea, o parâmetro a ajustado assumiria valores elevados, pois quanto maior este valor, mais abrupta é a modificação do metabolismo celular. Nos processos em questão, a variação da concentração de glicose na qual ocorre a modificação no metabolismo celular corresponde a um período de aproximadamente 3 horas, cerca de 15% do tempo total de cada batelada, mostrando a importância da consideração de transição gradativa entre as taxas de crescimento.

Com relação ao valor assumido da concentração de glicose que apresenta a variação máxima das taxas de crescimento, correspondente ao ponto de inflexão das taxas de crescimento em função da concentração de glicose, ocorre para concentração de glicose igual ao parâmetro b (G=0,925 g/L). Este valor de concentração pode ser considerado como sendo o instante pontual de transição do metabolismo celular, de forma a possibilitar uma comparação direta entre o modelo original e o modificado, levando a concluir que a levedura modifica seu metabolismo para concentrações de glicose entorno de 1g/L, e não para concentrações de glicose iguais a 0 g/L, como proposto inicialmente.

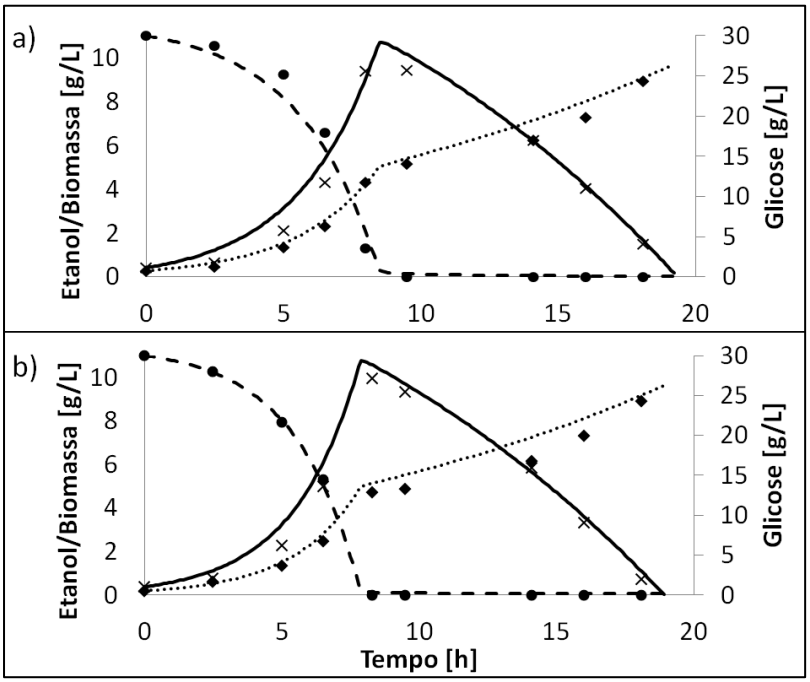

Figura 3.5: Simulação de dados *off-line* dos estados fermentativos (a) ensaio fermentativo 1 e (b) ensaio fermentativo 2, através de modelo dinâmico ajustado simultaneamente para ambos os ensaios fermentativos. (×) Etanol, (●) Glicose e (◊) Biomassa. As linhas representam os estados simulados pelo modelo.

A análise dos parâmetros a e b permite uma avaliação mais detalhada dos ensaios fermentativos, fornecendo índices capazes de caracterizar de forma quantitativa o metabolismo celular deste microrganismo. Modelos capazes de descrever transições no comportamento celular permitem o estudo de estratégias de operação otimizadas para o sistema em questão, elevando sua rentabilidade.

Os dados de Espectroscopia de Fluorescência 2D (EF-2D), fornecidos pelo equipamento Bio-View são estruturados na forma de tabelas bidimensionais, onde as linhas fazem referência aos comprimentos de onda de excitação e as colunas aos comprimentos de onda de emissão. A Figura 3.6 apresenta um diagrama genérico da forma como os dados de espectroscopia são apresentados, para cada uma das coletas de EF-2D. Neste diagrama é salientada a distinção entre os pares de fluorescência considerados válidos e inválidos.

A distinção entre os pares de fluorescência ocorre devido ao método de varredura realizado pelo equipamento. Uma vez que as medidas de fluorescência são realizadas variando os comprimentos de onda de excitação e emissão, ocorre que cada célula da matriz de dados é composta pela combinação entre um comprimento de onda de excitação e um de emissão, entretanto, como o fenômeno de fluorescência é devido à absorção e emissão de energia, o comprimento de onda emitido não pode apresentar energia maior que aquele usado para excitá-lo (como explicado no capítulo 2), desta forma, pares de fluorescência cujo comprimento de onda de emissão for menor que o de excitação, não apresentam informação real, apenas ruído de medida.

Devido a este fato, a matriz de dados de fluorescência, na forma como foi apresentada na Figura 3.6, possui zeros para valores de intensidade de fluorescência dos

pares acima da diagonal principal da matriz. Assim, os valores válidos de intensidade de fluorescência estão localizados abaixo da diagonal da matriz (pares de fluorescência situados na diagonal da matriz são aqueles cujos valores de comprimento de onda de excitação são iguais aos de emissão).

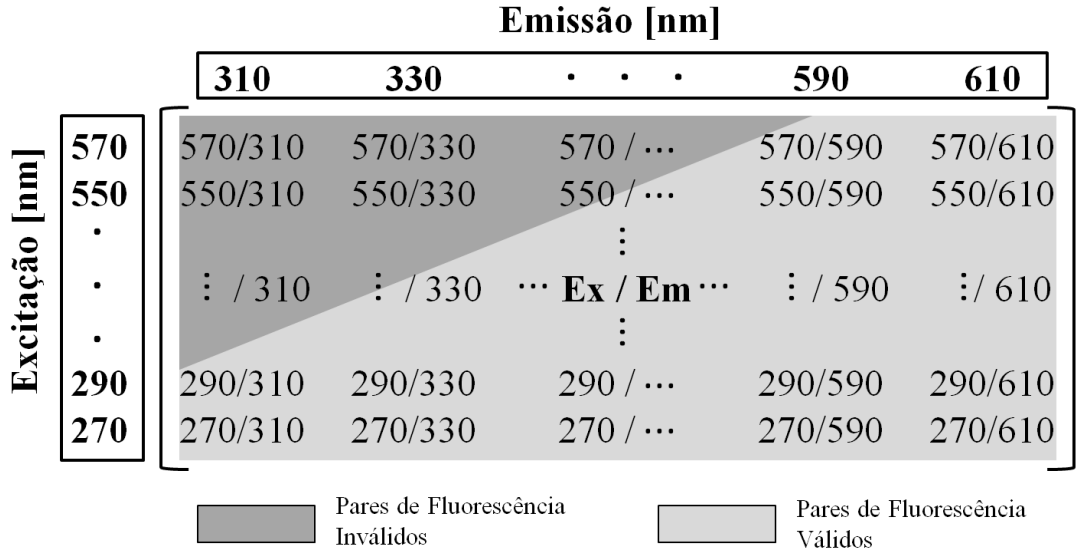

Figura 3.6: Diagrama genérico da forma como dados de Espectroscopia de Fluorescência 2D são apresentados, fazendo distinção entre pares de fluorescência válidos e inválidos.

As coletas dos espectros de fluorescência no decorrer de ambos os ensaios fermentativos foi feita com a mesma periodicidade. A Figura 3.7 apresenta três exemplos de espectros de fluorescência para cada um dos ensaios fermentativos, em tempos equivalentes.

A inspeção visual da Figura 3.7 mostra que, aparentemente, os ensaios fermentativos são equivalentes, evoluindo de forma similar durante o decorrer dos processos. A confirmação desta característica deve ser feita através de análises numéricas qualitativas, sendo imprescindível para avaliar a viabilidade de comparação entre os ensaios. Os dados experimentais não necessitam ser numericamente equivalentes, mas sim de forma qualitativa, sendo representativos do processo de forma geral.

De acordo com Kara *et al.* (Sep 2010), a diferença entre dois processos pode ser visualizada como a diferença entre os respectivos espectros iniciais de cada ensaio fermentativo. Associado a elevada complexidade de compostos presentes no meio reacional de processos fermentativos, e a sensibilidade do processo de Espectroscopia de Fluorescência, pequenas variações no meio reacional podem acarretar diferenças significativas nos espectros de fluorescência para os ensaios.

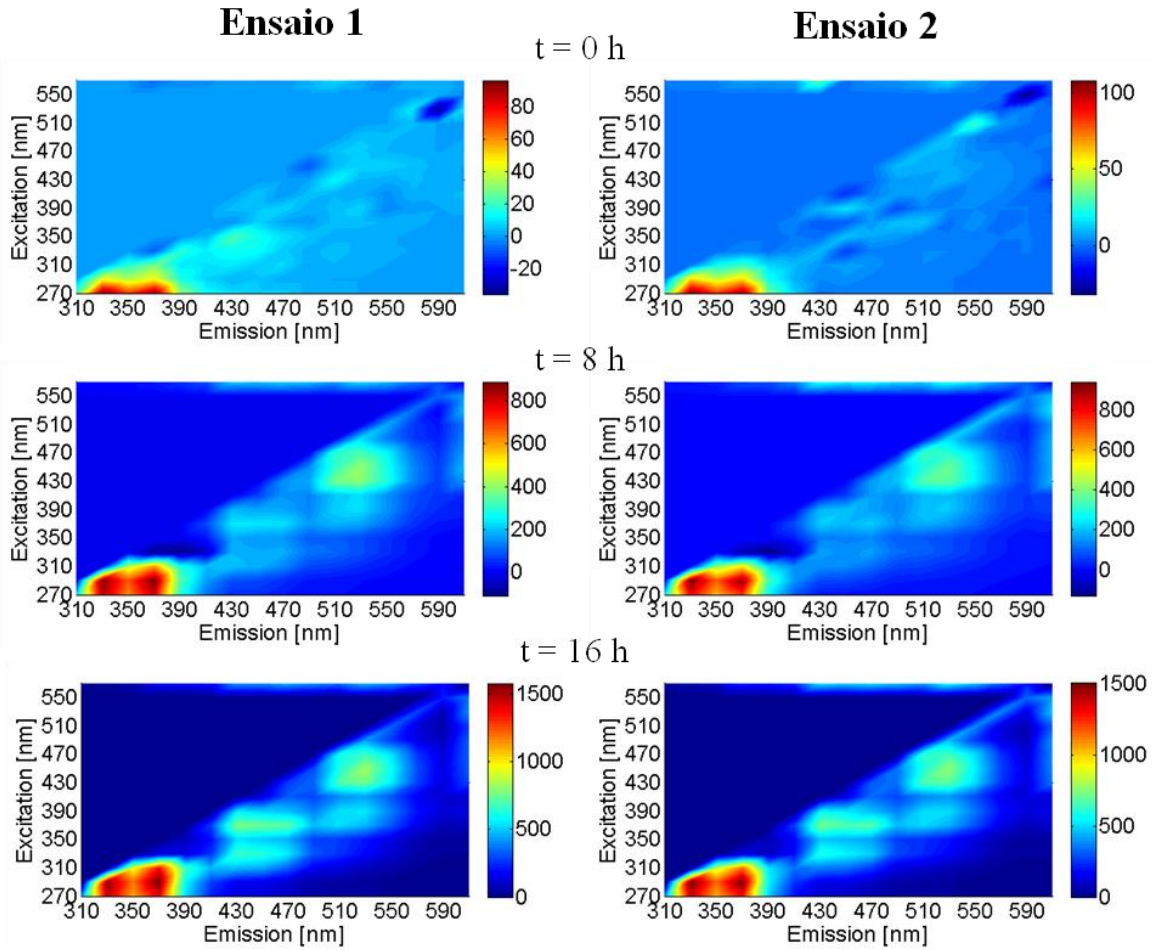

Figura 3.7: Dados de espectroscopia de fluorescência 2D para os ensaios fermentativos 1 e 2, coletados à 0 h, 8 h e 16 h a partir do início de cada bateladas.

Uma forma de avaliar a igualdade entre os meios reacionais é através da comparação entre os espectros de fluorescência no início do processo fermentativo, antes do microrganismo exercer influência no meio reacional. No instante de tempo igual a zero (*t=0* na Figura 3.7), o espectro obtido pode ser tratado como branco de amostra e se necessário, reduzido das medidas de espectroscopia subsequentes, na tentativa de retirar influências de variações do meio reacional na caracterização do processo. No conjunto de dados em questão, a comparação entre os espectros iniciais dos ensaios fermentativos mostrou que os meios são equivalentes, como apresentado pela Figura 3.8, onde é apresentado o módulo das diferenças numéricas, par-a-par, entre os espectros de fluorescência dos ensaios fermentativos.

A diferença entre os espectros iniciais dos ensaios fermentativos 1 e 2 apresenta uma única região com variações significativa na intensidade de fluorescência, compreendida entre [510 – 570] nm para excitação e [530 – 610]nm para emissão. As demais regiões não apresentaram diferenças significativas, mantendo a média do módulo de diferença em 4 unidades de intensidade de fluorescência.

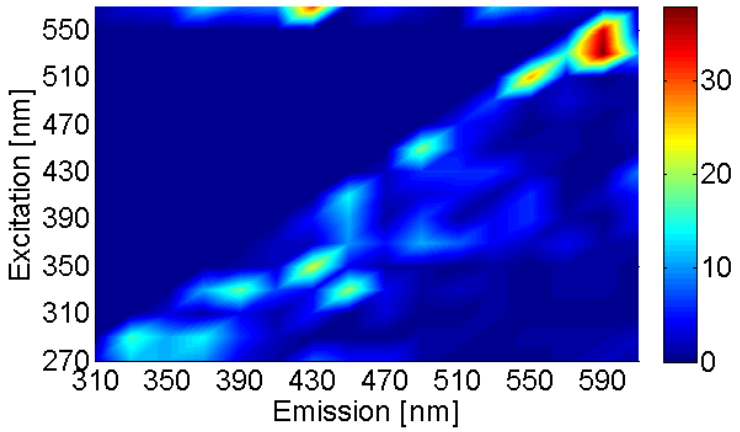

Figura 3.8: Módulo da diferença par-à-par entre os espectros de fluorescência, em t=0, dos ensaios fermentativos 1 e 2.

Uma avaliação mais clara da diferença real entre os espectros iniciais pode ser obtida através da comparação entre estes espectros após serem normalizados. Os espectros normalizados, utilizando a metodologia SNV, são novamente comparados e o módulo das diferenças, assim como os espectros normalizados, são apresentados na Figura 3.9.

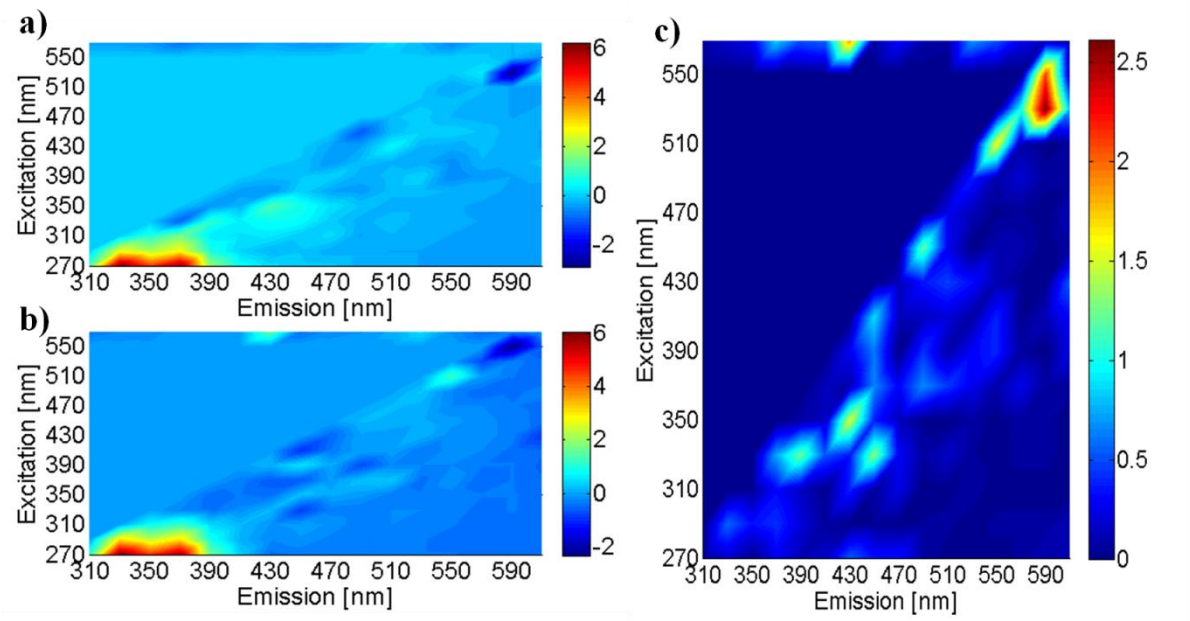

Figura 3.9: Espectros de fluorescência, em t=0, normalizados com método SNV para (a) ensaio fermentativo 1 e (b) ensaio fermentativo 2. (c) módulo das diferenças de intensidade de fluorescência, par-à-par, entre os espectros normalizados.

A comparação entre os resultados dos módulos das diferenças entre os dados espectrais iniciais dos ensaios fermentativos, normalizados (Figura 3.9c) e não normalizados (Figura 3.8), confirma a similaridade entre os dados de espectroscopia do meio reacional, indicando não ser necessário reduzir estes espectros dos dados subsequentes de cada ensaio. A principal diferença entre os resultados apresentados nas Figuras 3.8 e 3.9c diz respeito a diminuição da diferença da intensidade de fluorescência compreendida na região entre [270 – 310]nm para excitação e [310 – 390]nm para

emissão. Esta região é usualmente associada à fluorescência de fluoróforos como proteínas (ex: Tirosina, Triptofano e Fenilalanina) (Lindemann, *et al.*, 1998), sendo esperado que esta região apresente correlação com as variáveis de estado do processo em questão, já que estas moléculas estão diretamente associadas aos micro-organismos presentes no meio.

Apesar da equivalência entre os meios reacionais dos ensaios em questão, deve ser avaliada a viabilidade de comparação entre os processos ao longo de todo o tempo de fermentação. Bioprocessos são altamente suscetíveis a distúrbios e mesmo pequenas variações de temperatura nos processos, ou outras variáveis não monitoradas (cor do meio, concentrações de compostos secundários, etc.), podem ocasionar uma variação significativa na qualidade dos espectros de fluorescência (Lakowicz 2006).

A forma mais usual de avaliação qualitativa de dados de processo é através da análise de componentes principais. Aplicando PCA nos dois conjuntos de dados, previamente normalizados com SNV, e avaliando os *Scores* obtidos pode ser visualizado que os dados de fluorescência para os dois ensaios não apresentam diferenças que devam ser analisadas com maior aprofundamento. Os resultados da PCA dos dois ensaios fermentativos demonstram claramente o momento onde ocorre a transição de comportamento metabólico do microrganismo, permitindo sua determinação independente, sem necessidade de modelos específicos para tanto. A região onde ocorre a transição de comportamento metabólico esta localizada em PC1 equivalente à menos um e PC2 variando na faixa entre 2 e 4. A Figura 3.10 apresenta o gráfico dos escores dos componentes principais (PC) 1 e 2, para os dois ensaios fermentativos.

Uma vez que os pontos correspondentes aos dois ensaios assumem valores próximos e evolução equivalente no decorrer dos ensaios, pode ser concluído que os ensaios foram conduzidos da mesma forma, tratando assim de dados espectrais comparáveis do mesmo processo bio-químico. Caso houvesse uma diferença significativa entre os resultados obtidos por PCA para os dois processos, seria necessário promover diferentes pré-tratamento dos dados espectrais, visando reduzir influências causadas por variáveis externas não controladas (Kara, Anton *et al.* Sep 2010), entretanto, este não é o caso para o conjunto de dados em questão, sendo possível sua utilização, para fins quimiométricos, apenas utilizando normalização SNV.

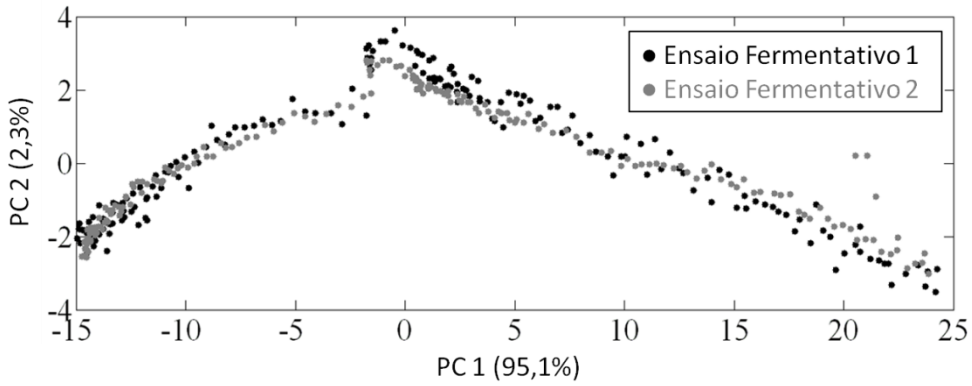

Figura 3.10: Componente principal 1 contra componente principal 2, para os dois ensaios fermentativos analisados.

3.2 Farinha de Trigo x NIR

Para avaliar a capacidade quimiométrica das metodologias propostas neste trabalho, em sistemas onde o objetivo é a caracterização de conjuntos de amostras aleatórias, foi utilizado como estudo de caso o conjunto de dados de espectroscopia no infravermelho próximo (NIR) de amostras de farinha. Todas as amostras foram paralelamente caracterizadas através de técnicas analíticas laboratoriais.

Farinhas de trigo são classificadas em diferentes tipos, dentro das classes comercializadas. O tipo de cada farinha especifica o conteúdo mineral da amostra, sendo expresso em miligramas para cada cem gramas de amostra (base seca), sendo aceita uma variação de 32%, para mais ou para menos, na quantidade de minerais. Assim, uma amostra de farinha de trigo classificada como sendo do tipo 1050 possui conteúdo mineral entre 910 – 1200mg / 100g.

Os tipos de farinha de trigo usualmente encontrados são: 405, 550, 812, 1050, 1600 e integral. Já farinha de centeio é encontrada usualmente nos tipos: 997, 1150 e integral, desta forma, a escolha do conjunto de amostras foi feita de forma a buscar a representatividade do universo de farinhas comercializados (Strohm 2012).

Ao todo, 34 amostras de farinhas foram coletas no comércio da cidade de Stuttgart, Alemanha. O conjunto de amostras é composto por farinha de trigo, farinha de trigo integral, farinha de centeio e farinha de centeio integral. A tabela 3.4 apresenta as amostras de farinha utilizadas, informando sua classificação, além da cidade onde são produzidas e seu nome comercial (marca).

Tabela 3.4: Conjunto de amostras de farinha.

Amostra	Classificação	Cidade	Marca
1	Trigo 405	Hamburg	Aurora Sonnenstern Mehl
2	Trigo 405	Neckarsulm	K-Classic Weizenmehl
3	Trigo 405	Heilbronn	Pizza Mehl
4	Trigo 405	Mannheim	Gold Puder Weizenmehl
5	Trigo 405	Landshut	Wiener Griessler
6	Trigo 405	Maxhütte	Qualitätsweizenmehl
7	Trigo 405	Hamburg	Gut&Gunstig Qualitatsweizenmehl
8	Trigo 405	Aalen	Weizenmehl
9	Trigo 550	Neckarsulm	K-Bio Weizenmehl
10	Trigo 550	Maxhütte	BioBio Weizenmehl
11	Trigo 550	Köln	ReweBio Weizenmehl
12	Trigo 550	Aalen	Weizenmehl
13	Trigo 550	Aalen	Weizenmehl S-fit
14	Trigo 550	Hamburg	Gut&Gerne Weizenmehl
15	Trigo 550	Mannheim	Gold Puder Weizenmehl
16	Trigo 1050	Neckarsulm	K-Bio Weizenmehl
17	Trigo 1050	Aalen	Weizenmehl
18	Trigo Integral	Neckarsulm	K-Bio Weizen Vollkornmehl

19	Trigo Integral	Bad Wimpfen	Kuchenmeister Weizen Vollkornmehl
20	Trigo Integral	Hamburg	Aurora Weizen Vollkornmehl
21	Trigo Integral	Köln	ReweBio Weizenvollkornmehl
22	Trigo Integral	Aalen	Weizenvollkornmehl
23	Centeio 997	Waldenbuch	Roggenmehl
24	Centeio 1150	Mannheim	Gold Puder Roggenmehl
25	Centeio 1150	Hamburg	Aurora Roggenmehl
26	Centeio 1150	Waldenbuch	Roggenmehl
27	Centeio 1150	Denkendorf	Roggenmehl
28	Centeio 1150	Altdorf	Bio-Roggenmehl
29	Centeio 1150	Altdorf	Roggenmehl
30	Centeio Integral	Bad Wimpfen	Kuchenmeister Roggenvollkornmehl
31	Centeio Integral	Denkendorf	Roggenvollkornmehl
32	Centeio Integral	Altdorf	Bio-Roggen Vollkornmehl
33	Centeio Integral	Altdorf	Roggen Vollkornmehl
34	Centeio Integral	Waldenbuch	Roggenvollkornmehl

As amostras de farinha foram caracterizadas analiticamente no Instituto *Fg. Prozessanalytik und Getreidetechnologie*, da Universität Hohenheim, em Stuttgart, Alemanha. Os equipamentos utilizados para a realização das análises quantitativas são apresentados na Tabela 3.5. Descrições detalhadas dos métodos analíticos, empregados na caracterização das amostras, podem ser obtidas em Strohm (2012).

Tabela 3.5: Análises laboratoriais realizadas nas amostras de farinha.

Análise	Marca e Modelo
Farinografia	*Brabender GmbH & Co. KG, Duisburg, Alemanha* *Modelo: FD0234H und Geratesoftware*
Extensográfica	*Brabender GmbH & Co. KG, Duisburg, Alemanha* *Modelo: EXEK/3 und Geratesoftware*
Amilográfica	*Brabender GmbH & Co. KG, Duisburg, Alemanha* *Modelo: Viskograph*
Dano por brotamento	*Perten Instruments AB; Kungens Kurga, Suécia* *Modelo: Falling Number 1400*
Teor de Glúten Úmido	*Perten Instruments AB; Kungens Kurga, Suécia* *Modelo: Glutomatic 2202*
Teor de Umidade	*Sartorius AG, Gottingen, Alemanha* *Modelo: MA51*
Determinação do Teor de Cinzas	*Heraeus Holding GmbH, Hanau, Alemanha* *Modelo: K1252*
Determinação do Teor de Proteínas	*Hach, Dusseldorf, Alemanha* *Modelo: DigesdahlR Digestions Apparatus*

Determinação do Teor de Proteínas	*GE Healthcare, Uppsala, Suécia Modelo: Spectro Photometer NovaSpecII*

Os dados de caracterização *off-line* das amostras de farinhas são apresentados na Tabela 3.6. Apenas variáveis mensuradas para todas as amostras são apresentadas nesta tabela. A tabela completa dos dados *off-line* pode ser obtida em Strohm (2012).

Tabela 3.6: Dados *off-line* do conjunto de amostras de farinhas.

Amostra	Umidade [%]	Tipo	Índice de Queda [s]	Proteína [g/100g]
1	12,06	522	296	9,3
2	12,7	497	337	11,3
3	11,62	504	328	10,9
4	12,48	506	300	9,3
5	11,99	371	361	8,5
6	12,87	525	318	8,6
7	11,84	574	365	9,6
8	11,88	635	360	9,8
9	12,24	642	335	8,9
10	11,26	1048	325	9,2
11	10,7	1782	324	8,3
12	10,51	1649	420	13,3
13	11,04	1515	279	11,5
14	10,85	1741	263	9,7
15	10,94	1260	176	9,2
16	10,61	1156	181	8,3
17	10,11	1559	215	8,2
18	12,45	538	308	10,3
19	12,20	486	352	11,2
20	11,17	635	378	11,1
21	10,73	614	324	12,2
22	10,85	507	375	10,2
23	11,70	626	359	11,9
24	10,24	1051	274	12,4
25	9,83	1655	304	11,6
26	10,91	1119	336	8,1
27	11,54	1382	255	8,8
28	12,40	1367	183	6,4
29	13,49	1584	226	6,9
30	12,93	1029	251	8,0
31	12,94	1234	221	8,8

32	13,72	1512	127	7,4
33	12,43	1182	179	8,0
34	13,13	1567	74	6,6

As amostras de farinha foram submetidas à análise de espectroscopia de infravermelho próximo (NIR) através do equipamento de análise multipropósito da empresa Bruker (Bruker Optik GmbH Ettlingen, Alemanha, Modelo: Multi Purpose NIR Analyzer). Frascos para contenção das amostras são completados com farinha e dispostos sobre o orifício de medição de espetro NIR refletivo no equipamento. Cada amostra é mensurada em triplicata, dando origem ao conjunto de dados espectroscópicos formados por 102 espectros. Entre cada medida os frascos são levemente agitados. A varredura dos espectros NIR coletados corresponde ao intervalo de número de onda entre 3600cm^{-1} e 12500cm^{-1}, equivalente ao intervalo 800nm à 2800nm para comprimento de onda, com incremento de onda variável ao longo do intervalo de varredura. A Figura 3.11 apresenta o equipamento utilizado na coleta dos espectros NIR, bem como o acondicionamento do recipiente para realização das medidas.

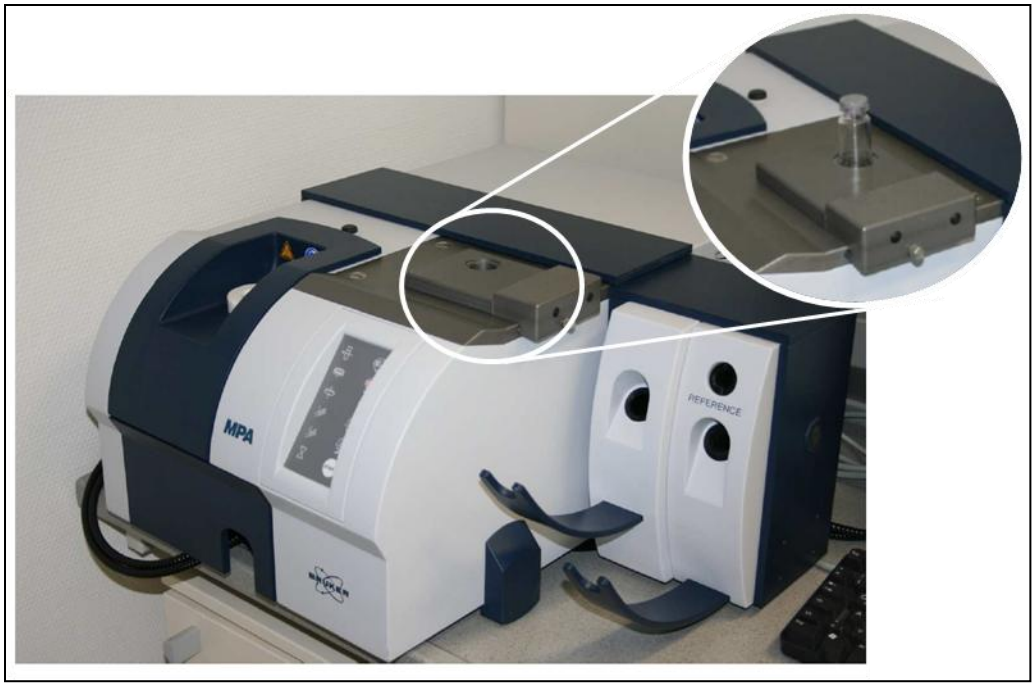

Figura 3.11: *Multi Purpose NIR Analyzer*, Bruker Optics GmbH, Ettlingen, Alemanha. No detalhe, recipiente de amostra acondicionado para medidas de espectroscopia NIR refletiva.

A avaliação qualitativa dos dados de espectroscopia NIR foi feita utilizando PCA. Nesta etapa, são utilizadas as médias das medidas em triplicatas de cada amostra, o que possibilita a análise da equivalência entre os grupos de amostras através das características dos espectros. A Figura 3.12 apresenta: (a) as médias das medidas NIR para as 34 amostras e (b) os dados de espectroscopia normalizados com SNV.

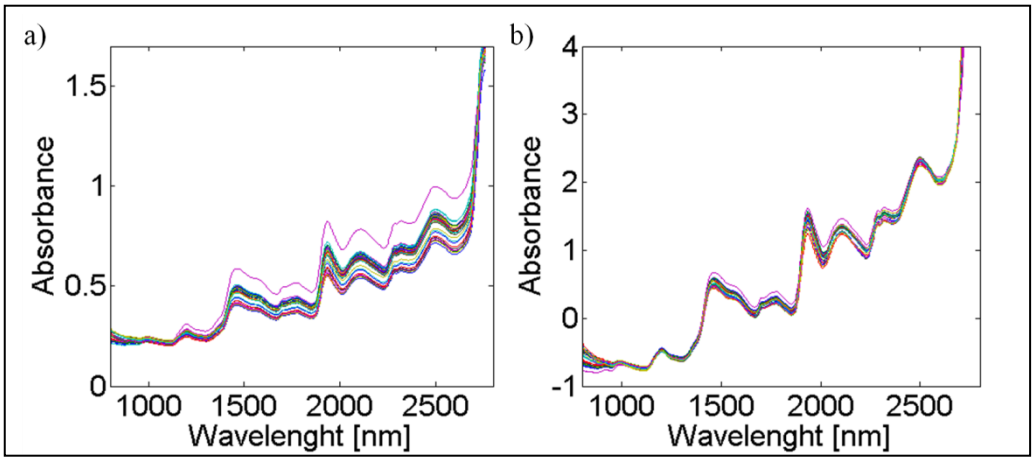

Figura 3.12: Medidas de espectroscopia NIR coletados para as 34 amostras de farinha. (a) Médias das triplicatas dos dados NIR originais e (b) dados NIR normalizados com SNV.

A avaliação qualitativa dos dados de espectroscopia das amostras é feita através da avaliação dos escores da análise PCA. A Figura 3.13 apresenta os gráficos dos primeiros cinco componentes principais apresentados pelas amostras de farinha. Nesta Figura são apresentados os PC's de dois à cinco em função do PC1, para cada amostra de farinha, indicando à qual amostra cada ponto se refere, além da distinção entre amostras de farinha de trigo ou centeio.

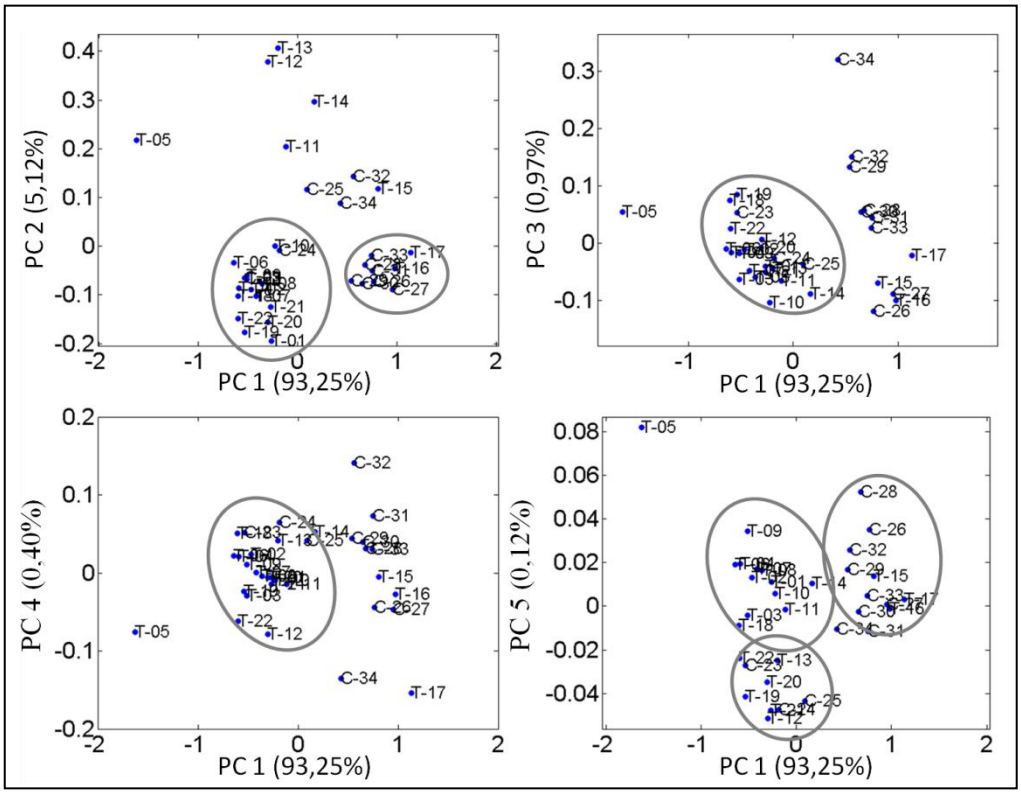

Figura 3.13: Gráficos de escores das amostras de farinha. T – Amostras de farinha de trigo. C – Amostras de farinha de centeio.

Os gráficos presentes na Figura 3.13 são usualmente utilizados para avaliação das amostras com relação à formação de clusters, ou distinção das amostras em grupos com qualidades de espectrais similares. No caso das amostras em questão, as amostras apresentam tendência de se agruparem de acordo com a origem das amostras de farinha, formando grupos de amostras correlacionadas estatisticamente. Apesar destes grupos poderem ser segmentados, sua distinção não é clara, mostrando que similaridades existem entre as amostras.

Apesar da equidade geral apresentada entre o conjunto de dados amostrais, a amostra número 5, correspondente a uma amostra de farinha de trigo, tipo 405 originária da cidade de Landshut, aparenta pronunciada diferença das demais amostras. Este resultado indica que esta amostra possui alguma diferença química com relação às demais amostras, não sendo representativa do restante do universo amostral disponível. Esta característica pode vir a resultar em resultados não satisfatórios para avaliações quimiométricas do conjunto amostral, entretanto foi optado por não retirá-la do conjunto amostral, com o intuito de avaliar o comportamento das técnicas propostas frente à existência de uma amostra não representativa do grupo.

3.3 PSCM - Quimiometria com ACO

Modelagem quimiométrica com componentes espectrais puros, ou PSCM (*Pure Spectral Chemometric Modeling*) é a metodologia para análise de espectros e obtenção de informação útil de processos, através da combinação da modelagem utilizando regressão multilinear e seleção de elementos utilizando métodos de otimização global.

Neste trabalho é proposta a ferramenta PSCM como uma alternativa aos métodos convencionais de análise qualitativa e quantitativa de dados espectroscópicos para o estudo e caracterização de processos químicos ou bioquímicos.

A metodologia PSCM é caracterizada pela utilização de modelos de regressão multilinear (MLR) associados ao método de otimização estocástica ACO (*Ant Colony Optimization*). Os elementos espectrais são filtrados, selecionando grupos de componentes espectrais com maior correlação com as variáveis de estado de interesse. Após seleção do grupo de elementos espectrais, estes são utilizados como variáveis de entrada em modelos multilineares, calibrados e então submetidos a testes na predição das variáveis calibradas.

A metodologia em questão é classificada como sendo do tipo assistida, pois para análise qualitativa e quantitativa dos espectros é necessário a utilização de dados amostrais previamente caracterizados.

Partindo de um conjunto de dados composto por amostras cujas variáveis de interesse estejam quantificadas, é realizada a análise de seus referidos dados espectrais para seleção do melhor conjunto de elementos capaz de descrever a variação destas variáveis dentro do conjunto amostral. Na Figura 3.14 são apresentadas as etapas do PSCM.

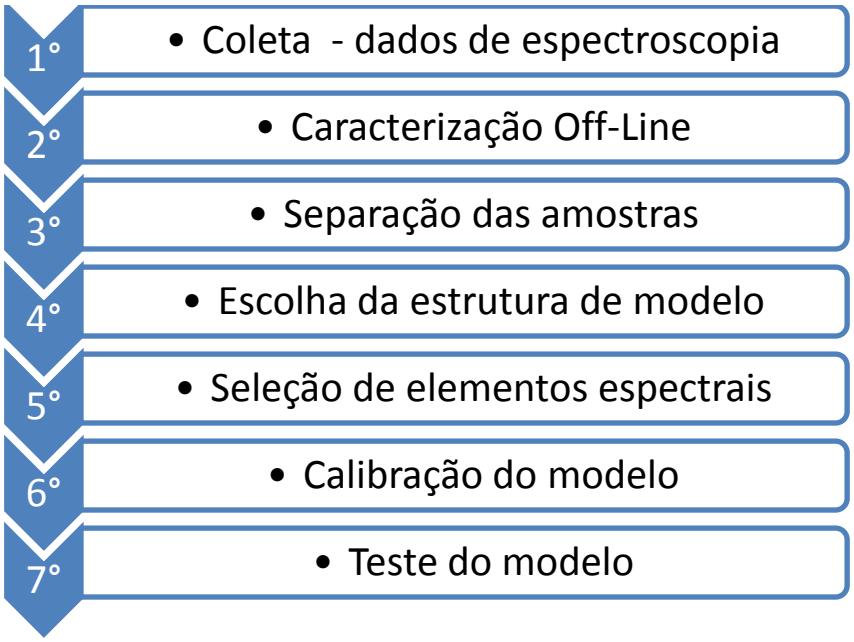

Figura 3.14: Representação esquemática da metodologia PSCM para estudo e caracterização de processos utilizando dados de espectroscopia.

As duas primeiras etapas compreendem a coleta e caracterização do conjunto amostral, selecionando as variáveis consideradas significativas para o processo e que necessitam de monitoramento. Nestas etapas estão subentendidas todas as fases de seleção do tipo de espectroscopia mais indicada para o processo, coleta e disposição das amostras, coleta de dados de espectroscopia, seleção e realização de ensaios laboratoriais, dentre outras.

A terceira etapa está relacionada com a escolha do método mais indicado para segmentação do conjunto amostral em dados utilizados para calibração e validação dos modelos quimiométricos. Nesta etapa, a característica do conjunto de dados amostrais é o fator significativo para decisão de qual a estratégia de validação cruzada (*Cross Validation*) a ser empregada. Para grandes conjuntos de dados, a metodologia *Y-rank* é a mais indicada, enquanto para conjuntos de dados com reduzido número de amostras, a estratégia *One Out* fornece resultados mais representativos.

Uma vez que ambos os conjuntos de dados utilizados são ricos em número de amostras, em ambos os casos, a estratégia *Y-rank* foi aplicada, resultando na utilização de 2/3 dos dados amostrais na fase de calibração e seleção de elementos espectrais, e 1/3 no teste dos modelos calibrados. Desta forma, a calibração de modelos, assim como a própria seleção dos componentes espectrais utilizados nestes, é feita independentemente das amostras utilizadas na validação, tornando as conclusões acerca da metodologia e aplicabilidade da técnica plausíveis.

A quarta etapa, por sua vez, trata da seleção prévia da estrutura de modelo que será utilizada na modelagem. Dentro da estrutura estão compreendidas características do modelo como número de variáveis de entrada e tipos de efeitos considerados. Nestas etapas do procedimento a utilização da regressão multilinear (MLR) propicia a vantagem do ajuste dos parâmetros do modelo ser feito através da resolução de um problema algébrico (capitulo 2.1.2), caso contrário, técnicas de ajuste de parâmetros através de métodos de otimização são necessárias, elevando o tempo computacional

significativamente. Modelos MLR permitem que efeitos quadráticos, iterativos ou de diferentes proporções sejam avaliados, desde que os modelos mantenham estrutura linear com relação aos seus parâmetros.

As etapas seis e sete correspondem à calibração e validação, respectivamente, dos modelos gerados pela combinação entre a estrutura de modelo e os elementos espectrais selecionados. Assim, estas etapas são a conclusão e avaliação do resultado das etapas anteriores, permitindo a análise qualitativa e quantitativa dos modelos e da capacidade de caracterização do processo estudado. Em ambas as etapas são feitos usos dos parâmetros estatísticos RMSE e R^2 para análise e comparação dos modelos. O núcleo da metodologia em questão é o método pelo qual o conjunto de elementos espectrais é selecionado na etapa 5. Desta forma, esta etapa é considerada a etapa determinante para o sucesso ou fracasso da modelagem do processo, onde é feita a seleção do grupo de elementos espectrais que melhor descreve as variáveis de estado analisadas, através da estrutura de modelo proposta na etapa 4. Nesta etapa, é avaliado o range espectral e selecionada a combinação de elementos ótima para predição das variáveis de estado selecionadas na etapa 2.

Uma vez que o número total de possíveis grupos de elementos espectrais é igual à combinação do universo de elementos disponíveis, combinados *n* a *n*, onde *n* representa o tamanho do grupo selecionado, a utilização de ferramentas especifica para este fim é obrigatória para viabilizar o procedimento.

A seleção do conjunto de elementos espectrais é feita através da metodologia ACO, onde o grupo de elementos é selecionado através da minimização da função objetivo que quantifica o somatório dos erros entre os dados medidos e preditos das amostras do grupo de calibração.

O algoritmo implementado neste trabalho é uma modificação da versão discreta do algoritmo apresentado por Dorigo e Gambardella (1997) para resolução do problema do Caixeiro Viajante (*Travelling Salesman Problem*), um problema de otimização combinatória de busca em um espaço de permutação.

O algoritmo de Dorigo e Gambardella é fundamentado na distribuição de *m* formigas em *n* cidades de forma a permitir que cada uma das formigas percorra um caminho fechado passando uma única vez por cada uma das cidades e percorrendo o menor percurso para isso. As duas características fundamentais deste algoritmo são a forma como é feita a seleção das cidades e a simulação da vaporização e reforço da marcação de feromônio.

A seleção das cidades é feita de forma que cada uma das formigas estando em uma determinada cidade, escolha a próxima cidade, excluindo de sua seleção todas as cidades já visitadas. Esta seleção é feita de forma que sejam levados em consideração dois fatores: um aleatório e um baseado na distribuição de feromônio nas trilhas entre as cidades.

A variação da intensidade de feromônio, por sua vez, é feita de forma que a intensidade de marcação de feromônio em todas as trilhas ligando as cidades passe por dois processos de modificação: (a) enfraquecimento da intensidade de marcação por vaporização, que ocorre através da multiplicação de um fator selecionado pelo usuário compreendido entre 0 e 1, após cada iteração, simulando a evaporação do marcador no

tempo, e (b) reforço da intensidade da marcação devido à utilização das trilhas pelas formigas.

A versão modificada, proposta neste trabalho, é baseada nos mesmos preceitos do algoritmo original, mas modificada para selecionar grupos de elementos espectrais, onde a ordem de seleção não influencia o resultado, ao contrário do problema original onde a ordem de seleção das cidades é o foco da resposta. Além desta diferença, a versão modificada também deve buscar um grupo seleto de variáveis dentro do espaço disponível, de forma que apenas as mais indicadas são selecionadas, descartando as variáveis menos importantes.

A nova versão, voltada à seleção de elementos espectrais, utiliza o chamado "vetor de feromônios" para armazenar a marcação associada a cada variável disponível. Este vetor unidimensional, de tamanho igual ao de elementos espectrais totais, faz comparação direta a matriz de feromônios do algoritmo de Dorigo e Gambardella, utilizada para seleção das cidades.

A Figura 3.15 apresenta um esquema do algoritmo implementado para resolução do problema de seleção de grupo de elementos espectrais utilizando ACO. O algoritmo está subdividido em quatro fases: inicialização das variáveis, inicialização da solução, rotina de busca e apresentação dos resultados.

Na fase inicial do algoritmo, são inicializadas as variáveis necessárias para dar início à resolução do problema de otimização. Nesta etapa são definidas ou carregadas no programa: (a) dados espectrais, dispostos na forma de matriz bidimensional, onde cada linha corresponde a uma amostra distinta e cada coluna corresponde a um determinado componente espectral, (b) vetor coluna de variáveis observadas, onde cada linha corresponde à respectiva amostra dos dados de espectroscopia carregados, (c) escolha do número de ciclos que o algoritmo realizará a busca pelo ótimo, (d) escolha do tamanho do exército de formigas utilizado na busca, (e) tamanho do modelo, indicando o tamanho do grupo de elementos espectrais a ser buscado, (f) tipo de modelo utilizado para avaliação do grupo de elementos espectrais, sendo por *default* do tipo linear sem interações, mas podendo assumir estrutura quadrática pura, quadrática com interações ou linear com interações, (g) valor inicial da trilha de feromônios, referente ao valor igualitário associado a cada elemento espectral antes do início do processo de otimização, de forma que todos os elementos iniciem com a mesma quantidade de marcador (valor *default* igual a 10^{-6}) e (h) taxa de evaporação de marcador, indicando o multiplicador do vetor de feromônios aplicado entre cada ciclo (*default* igual a 0,5, indicando que a cada ciclo todos os marcadores reduzem à metade de seu valor anterior).

Na segunda fase é feita a inicialização do vetor de soluções, através da resolução da função objetivo com a seleção aleatória de um conjunto de elementos espectrais. Esta etapa tem o objetivo de unicamente iniciar o vetor de soluções de forma a propiciar futuras comparações dos resultados contidos neste vetor com os próximos resultados obtidos da otimização, que serão gerados na próxima etapa. Até esta etapa do algoritmo, nenhum tipo de otimização foi realizado, apenas inicialização das variáveis.

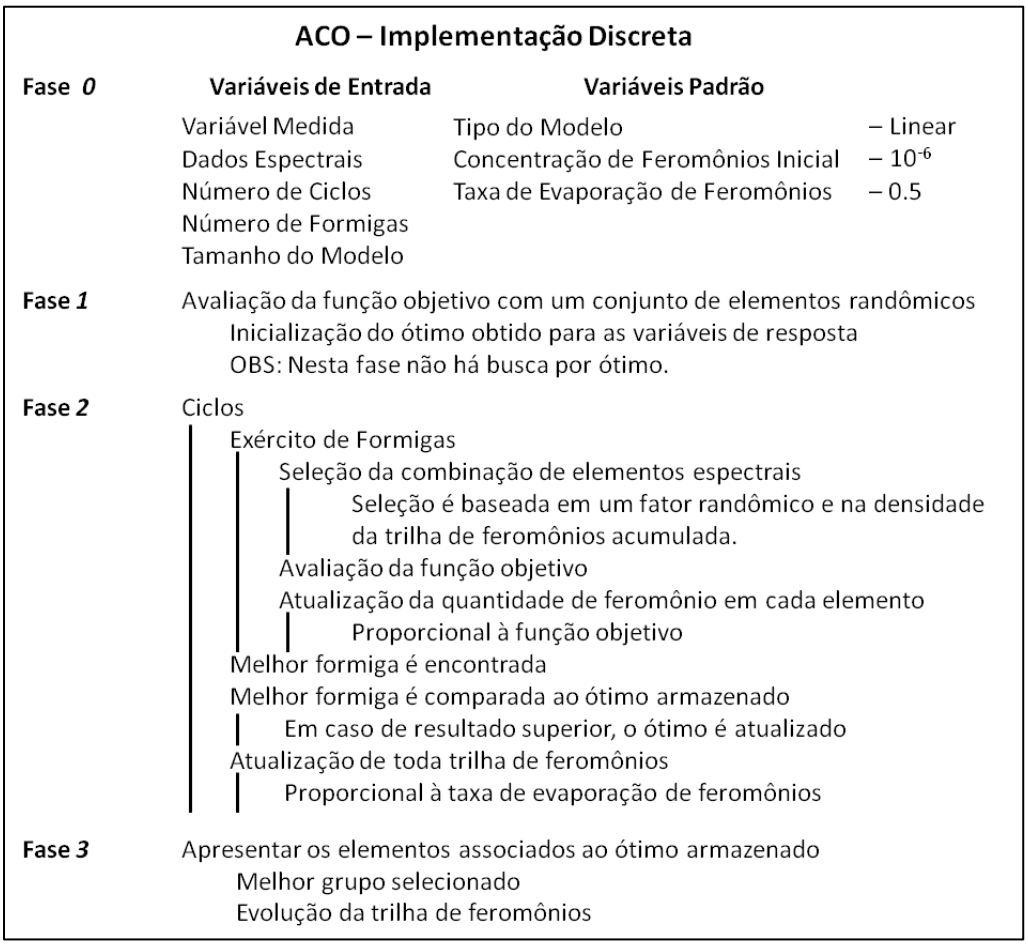

Figura 3.15: Representação esquemática do algoritmo ACO implementado para seleção de grupos de elementos espectrais.

O núcleo, propriamente dito, do algoritmo de otimização encontra-se na terceira fase de sua implementação, ou *Fase 2* (Figura 3.15). É durante esta etapa que o programa determina a melhor combinação de elementos espectrais para predizer a variável de estado observada, através da estrutura de modelo selecionada. Esta busca é realizada pela proposição de diferentes possíveis soluções (grupos de elementos) e comparados com o melhor resultado obtido até o momento (armazenado no vetor de soluções).

Nesta etapa, durante cada ciclo do algoritmo, o exército de formigas varre o universo de possíveis soluções na busca pelo grupo de elementos espectrais que forneça o menor valor para a função objetivo. A função objetivo, neste caso, trata do somatório da diferença elevada ao quadrado entre os valores observados da variável de estado e o predito pelo grupo de elementos selecionado, para cada amostra. Cada um dos indivíduos do exército de formigas escolhe uma grupo de elementos e submete ao teste da função objetivo. Caso o resultado seja inferior ao previamente armazenado, este é então substituído e o novo melhor resultado toma lugar no vetor de soluções.

A seleção de cada possível candidato a elemento espectral constituinte do grupo de predição, feita por cada formiga, é baseada em dois fatores, um aleatório e um baseado no vetor de marcadores (vetor da trilha de feromônios).

O fator aleatório garante que o algoritmo de busca não fique retido em possíveis mínimos locais, fazendo com que toda a região de busca seja avaliada. Este fator, na pratica, é implementado através de uma função que gera valores randômicos entre 0 e 1. Cada vez que um novo elemento deve ser selecionado e adicionado a um grupo por iniciar ou em processo de formação, o algoritmo aciona o "gatilho randômico" e utiliza seu resultado como fator de decisão para seleção do próximo elemento constituinte do grupo.

O valor fornecido pelo gatilho randômico é comparado com valor de densidade de feromônio acumulada (equação 3.10), de forma que o elemento que apresentar densidade de feromônio acumulada igual ao valor fornecido pelo gatilho randômico é inserido no grupo de elementos da solução.

A construção do grupo de elementos espectrais, realizada individualmente por cada formiga é feita ordenadamente, de forma que os elementos são escolhidos um a um. Cada elemento previamente escolhido por uma formiga é retirado do conjunto de possíveis opções para aquela formiga, durante aquele ciclo, retornando ao vetor de possíveis escolhas para as próximas formigas, e para ela mesma no ciclo seguinte.

Na estratégia original de Dorigo e Gambardella, a função de densidade de feromônios acumulada é obtida através do cálculo da densidade de feromônio relativa (equação 3.9), para cada elemento espectral em função do total de marcador presente nos elementos, que é então avaliada de forma cumulativa desde o primeiro até o último elemento espectral disponível (equação 3.10).

Nas equações 3.9 e 3.10, ρ_{Fi} indica a densidade de feromônio relativa do elemento espectral *i*, F_i indica a quantidade de marcador associado ao elemento espectral *i*, *N* indica o número de elementos espectrais constantes no vetor de elementos espectrais e C_{Fi} indica a densidade de feromônio acumulada desde o primeiro até o *i-ésimo* elemento espectral.

$$\rho_{Fi} = F_i / \sum_{i=1}^{N} F_i \tag{3.9}$$

$$C_{Fi} = \sum_{j=1}^{i} \rho_{Fj} \tag{3.10}$$

A Figura 3.16 apresenta um exemplo aleatório do procedimento de seleção de elementos espectrais através do gatilho randômico e uma trilha de feromônios simulada, composta por 1000 elementos espectrais, além de sua respectiva curva de densidade de feromônios acumulada. A partir do valor randômico 0,431 é feita a seleção do correspondente elemento espectral número 670, como demonstrado no exemplo da Figura.

O exemplo apresentado anteriormente demonstra a dinâmica de qualificação dos elementos espectrais, desempenhada no decorrer do processo de otimização, através da trilha de feromônios. A curva de C_F é sensível a valores elevados de marcadores, de forma que, elementos espectrais com altas concentrações de feromônio apresentarão taxa de variação da curva C_F, em função dos elementos espectrais, elevada, ao contrário de elementos com baixos valores de *F*, cuja taxa de variação da curva C_F é próxima de zero.

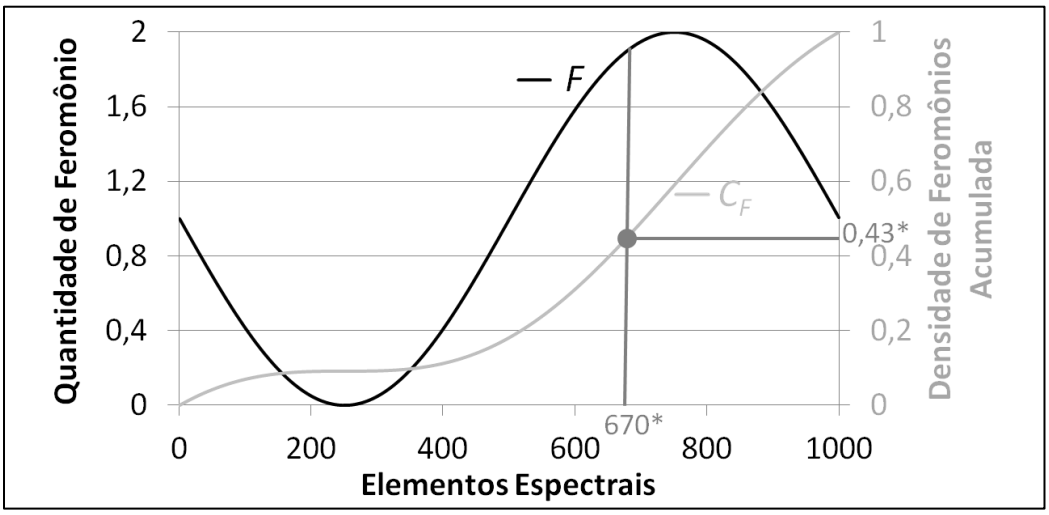

Figura 3.16: Exemplo de seleção de elementos espectrais utilizando a estratégia de Dorigo e Gambardella (1997). F_i corresponde à trilha de feromônios simulada e C_{Fi} à sua respectiva curva de densidade de feromônios acumulada. No exemplo, o gatilho randômico é acionado em 0,43, indicando a seleção do, respectivo, elemento espectral 670.

O fato de o gatilho randômico ser utilizado para seleção de elementos associados aos valores da curva C_F, e desta, por sua vez, apresentar variações significativas apenas para elementos com elevados valores de *F*, faz com que a probabilidade do algoritmo selecionar elementos com maior concentração de feromônios associadas seja elevada, priorizando assim a seleção dos elementos espectrais melhor avaliados pelo exército de formigas.

Com o intuito de melhorar o processo de avaliação e seleção dos elementos espectrais, foi promovida uma alteração no procedimento de cálculo da concentração de feromônios acumulada. A equação 3.13 apresenta a modificação proposta e utilizada nos algoritmos deste trabalho. Nesta nova estratégia, a importância atribuída aos elementos com valores elevados de *F* é maximizada frente os demais, fazendo com que elementos com maiores valores de *F* apresentem maior representatividade em C_{Fn}, aumentando sua probabilidade de seleção.

Testes desta nova abordagem foram feitos com conjuntos de dados para resolução do problema do caixeiro viajante e apresentaram resultados equivalentes aos obtidos pela implementação original, com diminuição do tempo computacional em cerca de 10%.

$$\boldsymbol{Fn_i = \tanh\left(\frac{3 * F_i}{max(F)}\right)} \tag{3.11}$$

$$\boldsymbol{\rho_{Fni} = Fn_i / \sum_{i=1}^{N} Fn_i} \tag{3.12}$$

$$\boldsymbol{C_{Fni} = \sum_{j=1}^{i} \rho_{Fnj}} \tag{3.13}$$

A utilização da função tangente hiperbólica, nos dados normalizados do vetor de trilha de feromônios, penaliza elementos com baixa quantidade de feromônio associada, aumentando a representatividade dos elementos com maiores concentrações de marcador na curva C_F. A Figura 3.17 apresenta as curvas de C_F e C_{Fn} para os dados simulados de *F* no exemplo da Figura 3.16. Apesar das curvas de densidade de feromônios

acumulada diferirem brevemente, para os dados do exemplo, pode ser confirmado que a representatividade dos elementos com baixas quantidades de feromônio foi reduzida, aumentando a probabilidade dos elementos melhor avaliados serem selecionados para compor os grupos de resposta.

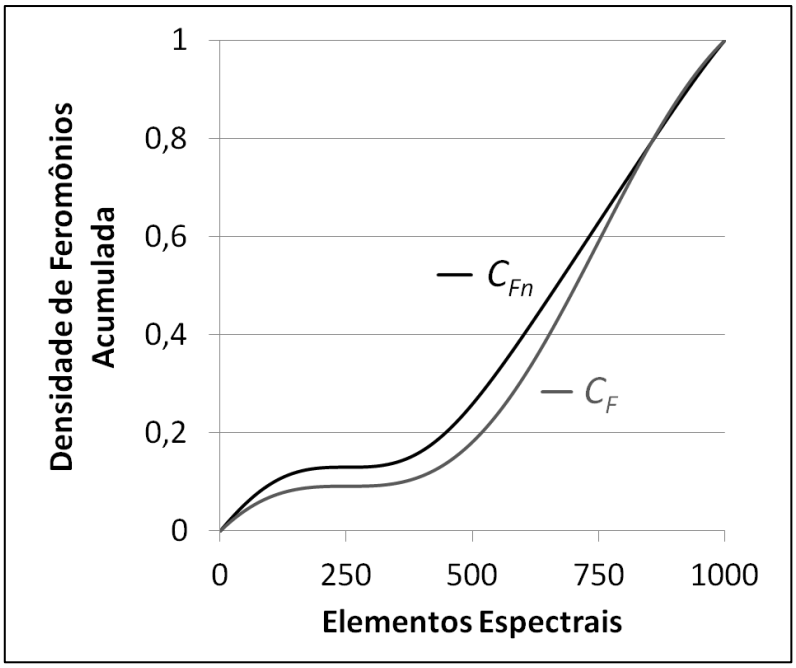

Figura 3.17: Curvas de densidade de feromônios acumulada para os dados do exemplo abordado na Figura 3.16, calculadas de acordo com a estratégia de Dorigo e Gambardella (1997) (C_F) e através da estratégia modificada (C_{Fn}), proposta neste trabalho.

Após cada formiga selecionar e testar um determinado grupo de elementos, a quantidade de feromônios associada a cada elemento deste grupo é atualizada, de forma inversamente proporcional ao somatório do erro quadrático entre os dados observados e preditos. Quanto maior o erro quadrático apresentado pelo modelo, menor é o incremento na quantidade de feromônios associada a cada elemento do grupo.

Esta estratégia faz com que elementos que apresentem maior correlação com a variável de interesse, ao serem adicionados a grupos de busca, produzam melhores resultados da função objetivo, aumentando o incremento da quantidade de feromônios nos elementos do grupo. Na evolução do processo de otimização, esta estratégia permite que os elementos espectrais sejam qualitativamente caracterizados em função de sua capacidade de predição da variável de interesse.

Ao final de cada ciclo do procedimento de otimização, todo vetor de quantidade de feromônios é multiplicado pelo fator que simula a evaporação da trilha, penalizando elementos que não foram selecionados pelo exército de indivíduos, no decorrer de cada ciclo.

No decorrer do processo de otimização, a trilha de feromônios evolui de forma aleatória nos ciclos iniciais, mas assumindo forma característica após certo número de ciclos. Ao final processo de otimização, a trilha de feromônios apresenta forma característica que depende unicamente do tipo de modelo proposto e dos respectivos dados de espectroscopia. Esta forma pode ser considerada como a “assinatura” dos

dados de espectroscopia associado à variável de estado, classificando de forma qualitativa as regiões do espectro em função da variável de estado.

A Figura 3.18 apresenta um exemplo meramente ilustrativo do processo evolutivo da trilha de feromônios no decorrer do procedimento de busca de grupos de elementos espectrais na descrição de um dado processo. No exemplo em questão, os dados espectrais são oriundos de espectroscopia RAMAN e avaliados para obtenção de grupos formados por cinco elementos, para predição de pH em amostras de carne de porco através de modelos lineares.

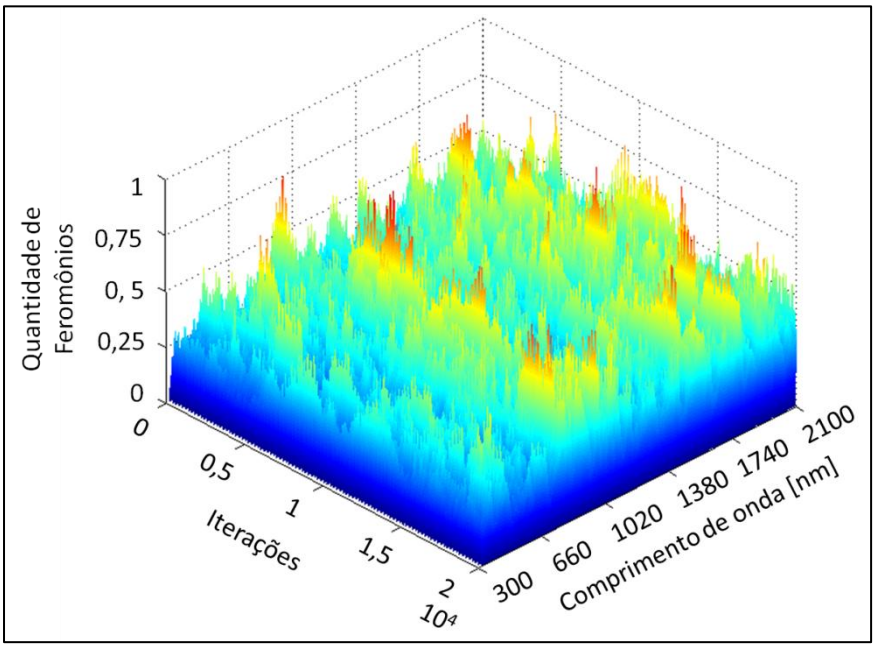

Figura 3.18: Exemplo ilustrativo da evolução da trilha de feromônios no decorrer da busca pelo conjunto de cinco elementos de espectroscopia RAMAN para predição de pH em amostras de carne de porco, usando modelos lineares.

Para finalizar o procedimento de seleção do grupo de elementos, na última etapa é apresentada a melhor solução obtida durante o processo de otimização. Este conjunto corresponde à melhor solução testada pelo exército de formigas no decorrer da rotina do algoritmo. Estes elementos, combinados através da estrutura de modelo proposta, são capazes de predizer os valores da variável de estado do processo apresentando o menor erro entre os dados preditos e medidos, sem a necessidade de manipulações nos dados de espectroscopia, ou mesmo coleta dos espectros completos.

Capítulo 4 – Wheat Flour Characterization using NIR and spectral filter based on Ant Colony Optimization

Abstract:[*] *The key objective for process optimization is to obtain higher productivity and profit in chemical or bio-chemical process. To achieve this, we must apply control techniques that closely correlate with our ability to characterize this process. Within this context, optical sensors associated with chemometrical modeling are considered a natural choice due to their low response time as well as their non-intrusive and high sensibility characteristics. Usually, chemometrical modeling is based on PCR (Principal Component Regression) and PLS (Partial Least Squares). However, since optical techniques are highly sensible and bio-chemical mediums are highly complex, these methodologies can be replaced by using chemometrical modeling based on Pure Spectra Components (PSCM). Our study applies PCR, PLS and PSCM for protein prediction in flour samples measured with Near Infrared Reflectance (NIR), comparing the three methodologies for on-line sensor project. We also outline the development of a spectral filter based on PSCM associated with Ant Colony Optimization. The results lead to our conclusion that the use of optical techniques works best when PSCM analysis is applied, as it allows the development of a spectral sensor for protein quantification in flour samples with less than twenty NIR wavelengths evaluated, selected from a total of 1150. The filtering tool showed favorable results in condensing relevant information from NIR spectral data, increasing R^2 from sample prediction by almost 60% for PCR models and 40% for PLS models, using 10% and 20% of full spectral data, confirming the viability of filtering methods.*

[*] *Published at Chemometrics and Intelligent Laboratory Systems*

4.1 Introduction

The ability to develop advanced control and optimization tools is intimately correlated with the ability to measure the state variables (Scheper, Hitzmann *et al*. 1999, Whitford and Julien 2007). Optical sensors are noninvasive, continuous and present low response time and cost with high sensitivity and resolution. More specifically, spectroscopy measurements - such as fluorescence spectroscopy, near infrared (NIR), multivariate FT-IR spectroscopy, Raman spectroscopy, and others (Clementschitsch, Jürgen *et al*. 2005, Rhee and Kang 2007, Whitford and Julien 2007, Roy and Pratim Roy 2009, Bosque-Sendra, Cuadros-Rodríguez *et al*. 2012) - allow us to detect several analytes simultaneously. All these features make optical sensors one of the most promising tools to be applied in chemical and biochemical processes (Ge, Kostov *et al*. 2005, Whitford and Julien 2007).

Spectral methods provide a very large amount of data that must be pre-processed to provide practical information for the user (Solle, Geissler *et al.* 2003, Kara, Anton *et al.* 2010, Warth, Rajkai *et al.* 2010). Therefore, the use of mathematical modeling is required in order to effectively measure analyte concentrations and/or material properties. As defined by Varmuza and Filzmoser (2008), "chemometrics concerns the extraction of relevant information from chemical data with mathematical and statistical tools". Successful methods to handle such data have been developed in the field of chemometrics: linear multivariate statistics such as multiple linear regression with factor analysis (FA-MLR), Stepwise Multi Linear Regression (Stepwise MLR), Partial Least Squares (PLS), Genetic Function Algorithm (GFA), Genetic PLS (G/PLS), Principal Component Analysis (PCA) or Principal Component Regression (PCR), as well as non-linear tools, such as Artificial Neural Network (ANN) (Clementschitsch, Jürgen *et al.* 2005, Rhee and Kang 2007, Roy and Pratim Roy 2009, Krishnan, Williams *et al.* 2011, Bosque-Sendra, Cuadros-Rodríguez *et al.* 2012, Farrés, Villagrasa *et al.* 2012). The most applicable methods are PCA, PCR and PLS, useful for quantitative analysis of spectroscopy data (Bro, van den Berg *et al.* 2002, Geladi, Sethson *et al.* 2004). These techniques are meant to provide a synthetic description of large data sets, allowing evaluations across the spectrum (Jolliffe 1986).

PCA is a powerful tool for data analysis, able to identify patterns in the data set and express data in a manner that highlights similarities and differences. Once patterns are found, the data set can be compressed without losing the main information. Several kinds of analyses use it to extract information related to physical and chemical properties from fluorescence matrices or for dimensionality reduction of fluorescence spectra in several systems (Tartakovsky, Lishman *et al.* 1996, Boehl, Solle *et al.* 2003, Guimet, Ferré *et al.* 2004, Rhee and Kang 2007, Kara, Anton *et al.* 2010).

PCR and PLS are commonly used with spectral data. After identifying the Principal Components, which account for most of the variance, these components can be used in regression. This method can transform highly correlated independent variables into uncorrelated Principal Components (PCs) (Rhee and Kang 2007). PCR has applications in Raman spectroscopy analysis and a few studies apply this method to analyze 2D fluorescence spectral data (Cooper 1999, Boehl, Solle *et al.* 2003, Otsuka 2004, Sorouraddin, Rashidi *et al.* 2005).

PLS, considered one of the most widely used multivariate calibration methods, is extensively applied to chemometric modeling of spectroscopic data to characterize biological process, such as in Infrared and 2D fluorescence spectroscopy, for instance. PLS

establishes mathematical relationships between on-line data (spectroscopy measurements) and off-line data (laboratorial analytical measurements) (Boehl, Solle *et al.* 2003, Eliasson Lantz, Jørgensen *et al.* 2006, Surribas, Geissler *et al.* 2006). Despite all these applications of PCA, PCR and PLS, many authors have produced comparative analyses for these methods in real data process. One example is the work of Rhee *et al.* (2007), where the authors compared PCR and PLS for characterization of fermentations using 2D fluorescence spectroscopy.

In their research, they analyzed two aspects not usually explored in literature: (i) the use of pure spectra for chemometric modeling without the need for statistical pretreatment of data sets and (ii) extension of this methodology for creation of a more robust procedure for on-line sensor development. In this sense, Shamsipur *et al.* (2009), Hemmateenejad *et al.* (2011) and Allegrini and Olivieri (2011) focused their attention on the use of Ant Colony Optimization (ACO) as a tool for wavelength selection, comparing the different possible implementations of ACO with other heuristic optimization methodologies, such as with genetic algorithms. However, a comparison between the well-established methods for spectral analysis - PCR and PLS - with the methodology using ACO, highlighting the advantages and disadvantages of each method is missing.

Thus, the main objective of our study is to compare well-known chemometric PCR and PLS methods with a non-trivial approach, which uses models directly based on spectrum components for state variable prediction. This method improvement is called Chemometric Modeling Based on Pure Spectra Components (PSCM) and uses the ACO as a tool for spectral component selection and chemometrical modeling.

We use a set of flour samples characterized using NIR spectroscopy to compare the chemometrical methods used. Given that the evaluation of NIR spectroscopy with PLS is considered a standard methodology for characterizing flour (Cocchi, Corbellini *et al.* 2005, Ait Kaddour and Cuq 2009, Li Vigni, Durante *et al.* 2009, Vigni, Baschieri *et al.* 2011), the use of this specific data set are ideal to compare the results of chemometric modeling using PSCM with standard methodologies.

4.2 Methodology

4.2.1 Experimental Data Set

The experimental data used in this work includes 34 samples of different brands and kinds of flour measured in triplicate to determine several important variables. In our study, however, we will only take the protein content values into consideration. All the samples were off-line and characterized through a farinograph analysis (Brabender GmbH & Co. KG, Duisburg, Germany, model FD0234H). We evaluated the protein content with a Digestion Apparatus (Digesdahl® Hach - Düsseldorf, Germany). Parallel to the farinograph analysis, we characterized the samples with NIR spectroscopy measurements. We performed NIR measurements in a Multi-Purpose NIR Analyzer (Bruker Optik GmbH - Ettlingen, Germany), varying wavelength from 800 nm to 2800 nm.

Figure 4.1 shows the average of protein concentration for data set samples. The off-line protein data set presents the segmentation of flour samples into calibration and test groups, which is necessary for chemometrical analysis.

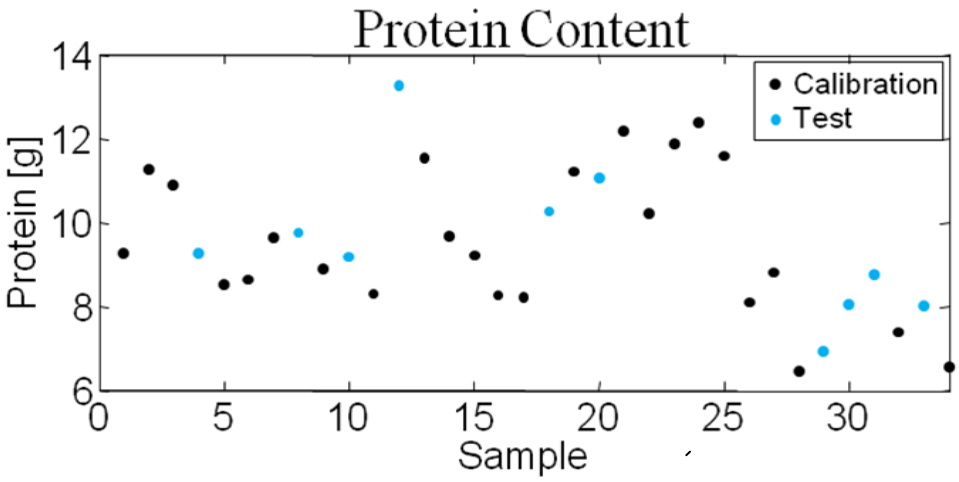

Figure 4.1: Protein content in sample set and segmentation in calibration and test set.

We chose this data set based on specific characteristics, such as protein content variability and spectral range. The sample's spectral data are composed by a large number of wavelengths that fall between 800nm and 2700nm, with a variable interval, totalizing 1150 spectral NIR elements.

4.2.2 Chemometric Methods

All implementations and calculation routines were made with MATLAB software (Ver. 5.3.0.10183 R11, The Mathworks, Inc., Natick, USA) and used 30 minutes maximum for running.

Regardless of which chemometric method is applied, we calibrate and test them all in different sample groups. Calibration and test groups are determined by selecting samples with protein contents that represent the full range inside each sample group. Figure 4.1 shows the segmentation used.

***Principal Component Analysis – PCA*:** To apply this method, the data set should be structured as a single matrix, with samples listed in rows, and individual spectrum intensity values listed in their respective columns (Jolliffe 1986).

PCA allows the reduction of dimensionality of multivariate data to *n* principal components (PCs), decomposing a spectral data matrix as a sum of outer product of vectors plus a residual matrix. Each PC captures the largest possible amount of variance remaining at the step. Therefore, PCA adequately describes the spectral data using far fewer factors than the original number of components (columns) in the original matrix (Rhee and Kang 2007).

For qualitative PCA analysis, we use triplicate measurements of NIR spectroscopy without mean calculus, allowing a qualitative comparison between samples and NIR reproductive measurements.

Principal Component Regression – PCR: It is not a methodology applied directly to the original spectral data, but is correlated and applied to the PCs obtained by PCA. Given

that PCs are mutually orthogonal, it is possible to avoid the typical problem of collinearity and high correlation which arises in many regression techniques. In this methodology, the PCs are combined in a way to predict the output data matrix, using multivariable linear regression (Liu, Kuang *et al.* 2003).

The procedure for PCR is divided in three steps: evaluation of PCA and determination of more significant PCs, evaluation of ordinary least squares regression on the selected components and computation of model parameters for the selected explanatory variables (Upton, Cook *et al.* 2008).

Partial Least Squares – PLS: This is a multivariate statistical technique with the goal of correlating two data sets and making a prediction of one set from another. It is meant to identify the factors (latent variables, LVs) that not only capture the largest amount of variance in data, but also provide a linear correlation between the spectral data and process variables.

The inner relation in PLS can be expressed as a sum of the outer product of score matrix and identity matrix plus a residual matrix. The regression equation of output data matrix can be written as the sum of the product of spectral data matrix and regression coefficients related to the weights, which can be considered additional loading matrices that express the correlation between spectral data matrix and output data matrix (Geladi and Kowalski 1986).

Pure Spectra Chemometric Modeling – PSCM: This chemometric analysis has two main pillars: the selection of pure spectral elements and model adjustment for state variable prediction.

Figure 4.2 illustrates the steps for chemometric modeling using the PSCM methodology. The full methodology is divided into blocks in way to present some characteristics about each step.

The selection of the spectral group aims to choose spectrum components that present direct correlation with state variables, discarding possible noise and errors caused by spectra regions not significantly related to interest variable (Skoog, Holler *et al.* 2007).

Selected group of spectral variables are used as input data in multi linear models with one output (MISO models - Multiple Input Single Output). Those models are linear in relation to their parameters, allowing calibration by analytical least squares solution.

The use of multilinear models is not mandatory, although their parameter calibration allows for an algebraic resolution, significantly decreasing the time needed for adjustment of models as well as for spectral group selection. Any model structure can be applied, depending solely on the kind of spectral data available and process in the study case.

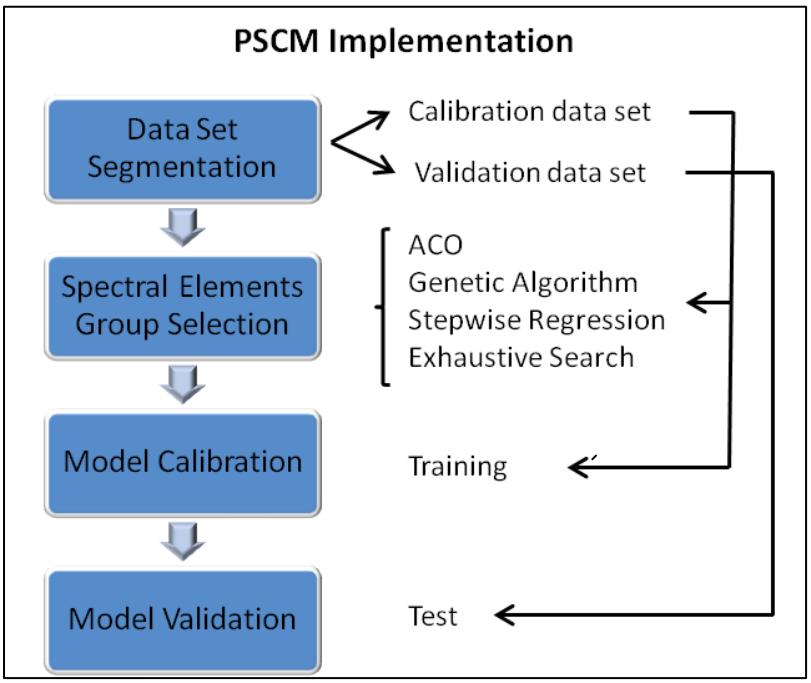

Figure 4.2: Schematic PSCM implementation.

The PSCM methodology requires previously characterized data (spectroscopic and laboratorial), providing relevant process information about spectral behavior as function of process variables. The larger the experimental data set for models proposed, the more robust and reliable the models obtained will be. The representativeness of process characteristics in the experimental data set is crucial to identify qualitative spectral information and consequent chemometrical model development.

The search for independent variables in PSCM is crucial for method efficiency, and can be done through many different approaches. The simplest way applies an exhaustive search, where all the possible combinations of spectral variables are tested and the best one is selected. Although this technique ensures that the optimal one is found, depending on the number of spectral variables and model size, the number of possible combinations can be so high that computation time makes it impracticable.

To solve this problem, we apply optimization techniques to select the best spectral component set to describe a particular process variable. The benefits obtained from this methodology are not only the stability of the model in terms of collinearity in multivariate spectra, but also the interpretability of the relationship between the model and the sample compositions, as initially presented by Allegrini and Olivieri (2011). Previous studies in our group tested several methods for spectral variables selections (Ranzan, Trierweiler *et al.* 2012, Masiero, Trierweiler *et al.* 2013, Ranzan, Ranzan *et al.* 2013), and despite PSCM allow the use of any variable selection method, functionalities of ACO (e.g. easy implementation, time consuming, qualitative spectrum data, etc) led us to decide for this method.

ACO is a methodology meant to solve minor optimization issues (e.g. travelling salesman problem). This optimization tool is based on the behavior of real ants; more specifically, by the indirect communication between them within the colony using chemical pheromone secretion (Dorigo and Blum 2005, Dorigo, Birattari *et al.* 2006,

Mullen, Monekosso *et al.* 2009). Figure 4.3 presents a schematic algorithm for the discrete version of ACO implementation. The main idea behind this is that, in real ants, the convergence of ant trails toward the shortest route between the food source and the nest is a result of the tendency of ants to follow a trail that contains a higher concentration of pheromone deposit (Deneubourg, Aron *et al.* 1986).

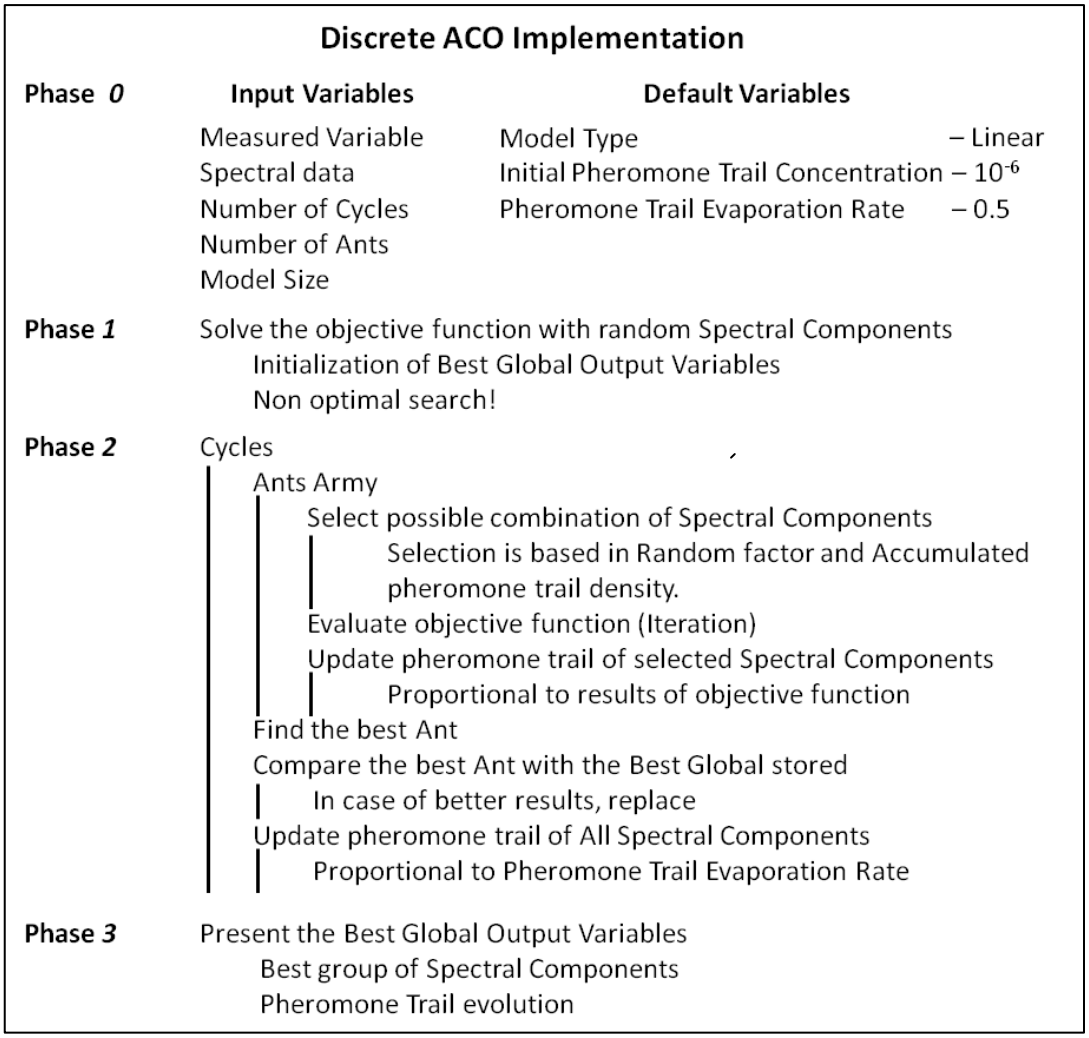

Figure 4.3: Schematic algorithm of Ant Colony Optimization.

Our implementation of ACO in this study is a discrete version, where the selection of spectrum components is made using a random factor associated with a pheromone density function. Pheromone is added to spectrum components by ants as a function of residual error between state variable prediction and measurement. More details about Ant Colony Optimization algorithms can be obtained in Mullen *et al.* (2009).

We analyzed proposed models according to two statistical parameters for accuracy ranking (calibration and test phases use the same statistical analysis). These parameters are RMSE (Root-Mean-Square Error) and R^2 (Coefficient of Determination), calculated according to equations 2.12 and 2.13, respectively. RMSE is named differentially when calculated in the calibration or prediction phase: RMSEC for calibration and RMSEP for prediction.

Spectral data Filter: PSCM analysis using the ACO algorithm presents the advantage of providing qualitative information about spectral elements as a function of interest variables.

Pheromone trail (internal vector of ACO) works as independent numerical information which ants use to probabilistically construct solutions to the problem where ants adapt during the algorithm's application to reflect their search experience. This vector is updated after each interaction of ACO and at the end of the optimization routine. It contains information on each spectrum component's importance, and, more importantly, the significant regions for variable prediction.

Allegrini and Olivieri (2011) previously described this characteristic of ACO, analyzing the advantages of ant colony optimization use combined with Monte Carlo repeated calculations to discard irrelevant spectral regions when PLS regression analysis is performed on NIR spectroscopy data of sugar cane, corn, octane and synthetic samples.

The methodology applied is based on interest variable spectral signature, captured by PSCM modeling. This information is used to filter the total amount of spectral data, selecting the more significant spectral elements from the spectral range associated with interest variable, based only on pheromone trails. Thus, we discard irrelevant spectral elements, reducing the amount of noise and non-significant information added to PCA, PLS or other chemometric methodologies.

Since the pheromone signature presented by interest variables is a function of the model structure evaluated during PSCM, an average pheromone trail is obtained by taking the mean of pheromone trail vectors obtained using different model structures.

All structure models tested on PSCM methodology are multi-linear, so filtering viability is tested using mean pheromone trails of PSCM models from one to ten input variables and ordered to select spectral components.

Filtered spectral data is tested using PCR and PLS chemometrical models, comparing protein prediction results using full and filtered spectral data.

The filtering method viability is tested using the sample set previously presented, as well as the triplicate measures without mean spectra, allowing larger PCR/PLS models, given that the intention is to test the filtering tool, not the technique's prediction capability.

4.3 Results

4.3.1 Process Characterization

The application of PCA starts with the appropriate scaling of spectral data. Previously tests showed that for the present data set, mean centering normalization provides better results than non-normalized data, fact that implies on the use of normalized data chemometric analyses. This procedure is not mandatory, once each case should be seen individually (Gaydou, Kister *et al.* 2011). This procedure is completed in each spectral

column, reducing each value from the column mean and dividing by the standard deviation of columns.

Figure 4.4 presents the normalized variance of principal components obtained in PCA and the scores from the first and second principal components. The graphic of variances (figure 4.4(a)) shows that the sum of variance of the three first PC's represents around 82% of all data variance. We used 102 NIR measurements (triplicate measurements for 34 flour samples) for this first analysis.

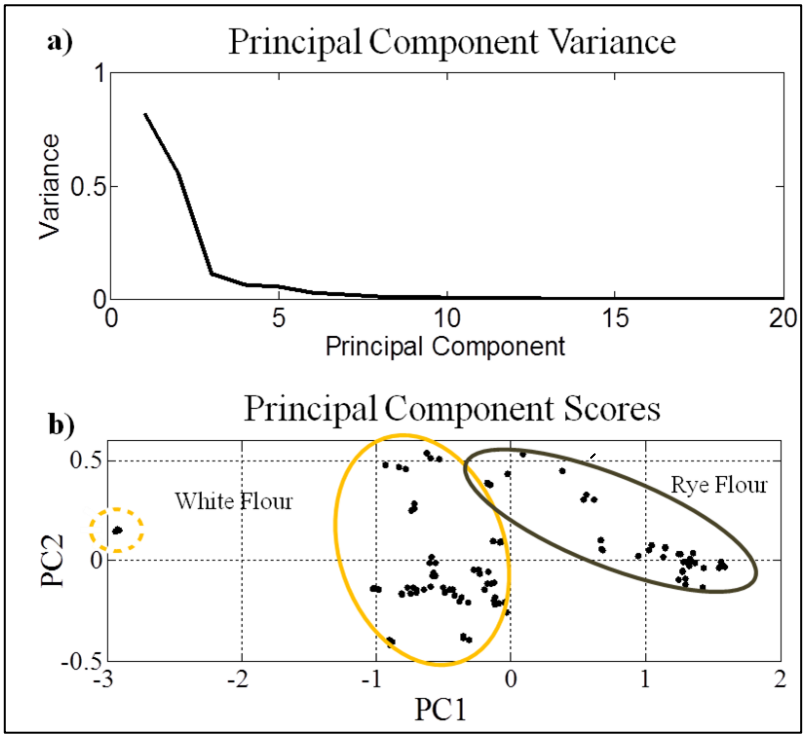

Figure 4.4: (a) PCA results with variance (only 20 more significant principal components). (b) Scores plotting from NIR spectroscopy of triplicate flour samples.

The score plot (figure 4.4(b)) shows that data can be separated into two big groups according to sample nature (rye flour and white flour), indicating the high correlation and similarity between the samples. Figure 4(b) also shows the high reproducibility of NIR measurements grouped by samples once triplicate measurements presented equivalent score plots. Despite generally qualitative similarities presented by flour samples, Sample 5 (hatched border) presented significant difference from the two distinguished groups. We did not, however, remove it from the data set.

The difference in spectral data presented by Sample 5 cannot be explained by experimental error in NIR spectroscopy measurements, once all triplicate measurements presented good reproducibility. In this case, the probable explanation is that this sample has significant chemical differences in comparison to the other samples, a fact that does not influence chemometric models comparison.

All the models have linear structures, not including interaction or quadratic effects of independent variables in the prediction of dependent variables. Many different models are fitted for the three methodologies with different sizes, using 1 to 20 independent

variables. We determined this huge interval to evaluate the behavior of prediction model capability and determine which methodology is better to run a chemometric analysis.

In PSCM the selection of model size is done by selecting the number of spectral components that will be used as independent variables in the model. It involves a significant broadening of search universe for the best group of elements to be used.

Given that PSCM has no pre-treatment of spectral data of any kind and input variables are the pure spectral components, the issue becomes choosing a subset of spectral channels that result in small prediction errors.

Figure 4.5 presents the results obtained in stages of calibration and prediction test of chemometric models using PCR, PLS and PSCM. Analyzing the values of RMSE and R^2 for PCR and PLS, we can conclude that the prediction of protein content in the prediction stage is more accurate in PCR for models with only one input variable; as for PLS, the best result is obtained for models with seven input variables. As for model size, PLS presents better results for models with less than seven input variables; however, there is an inversion for bigger models and PCR presents better results.

PSCM results presented better values of R^2 and RMSEP, when compared with standard chemometric models. For almost all sizes of multi-linear models analyzed, PSCM presented the best results as well as the best models (twelve and twenty input variables) for prediction of protein content using NIR data.

Comparing the three techniques, we can draw the conclusion that models obtained by PSCM present significant differences in prediction capability using a reduced number of independent variables. Analyzing the increment in accuracy when increasing the model size, models obtained through PSCM with twelve and twenty independent variables (R^2 for validation equal to 0.9) presented the best results for protein prediction. This shows that models based exclusively in pure spectral components could predict protein content in flour samples using less information than standard PCR or PLS models for the same amount of input variables.

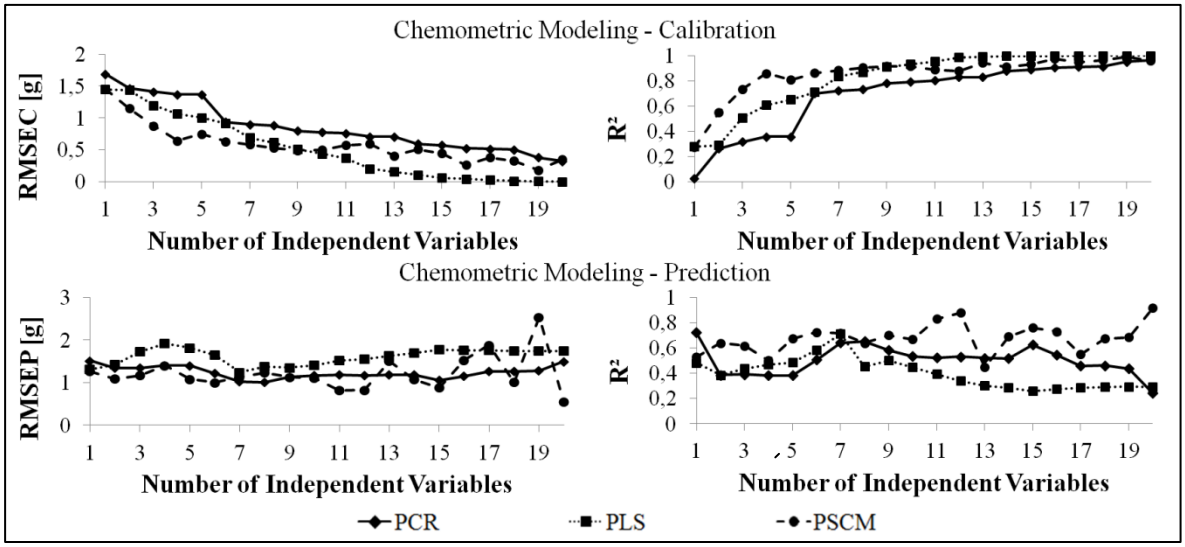

Figure 4.5: Effect of independent variables in R^2 and RMSE values for PCR, PLS and PSCM (calibration and validation) for samples of flour characterized by NIR spectroscopy.

The objective in variable selection is to reach the expected state variable prediction without the use of intuition or complementary selection methods. As exemplified with the experimental data set, the modeling process using the spectrum components results in better models with less independent variables because those components are directly correlated with the interest variable and, for the most part, provides useful information.

As the PSCM selected components in different regions of the spectrum, we can consider protein content in flour to be associated with three or four spectral regions. We base this analysis on the increment of chemometrical model size in PSCM with the increment in the accuracy of protein content prediction. Basically, using four spectrum components in PSCM models, we can achieve a R^2 around 0.8. When we add more input variables in models, however, this R^2 has no significant increase, indicating that these new input variables are providing redundant information.

4.3.2 Spectral Data Filter

Figure 4.6(a) shows the evolution of a pheromone trail during one PSCM routine for protein content prediction. The algorithm performs 200 cycles with an army composed of 100 ants and searches for a combination of four spectral components. Figure 4.6 also presents the pheromone trail mean (Figure 4.6(b)) during the routine, highlighting the region with high pheromone concentration.

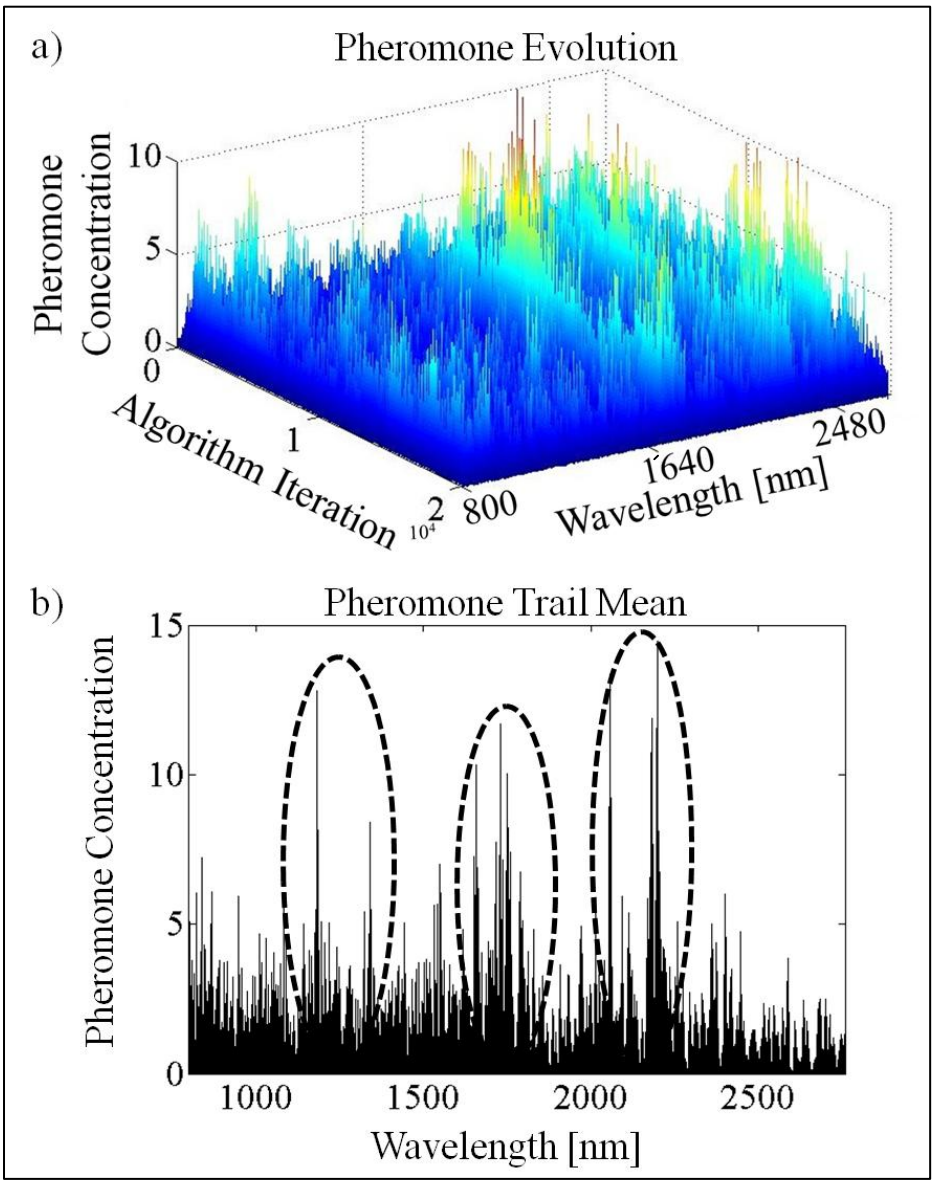

Figure 4.6: (a) Pheromone trail concentration (dimensionless) evolution in search for NIR spectra region to protein prediction and (b) pheromone mean values during optimization with indication of significant spectral regions for protein content in flour.

We compare pheromone trail highlights with our theoretical knowledge on flour characterization using NIR spectroscopy. Since NIR spectroscopy is a standard methodology for flour analysis (Technology 1986, Technology 1995), some molecules' spectral overtones had already been mapped (Wang 2010). This information can be compared with the results on pheromone spectral highlights.

Figure 4.7 presents an NIR reflectance spectrum of typical bread wheat flour with the indication of functional groups associated with some characteristic flour overtones (Sun 2008).

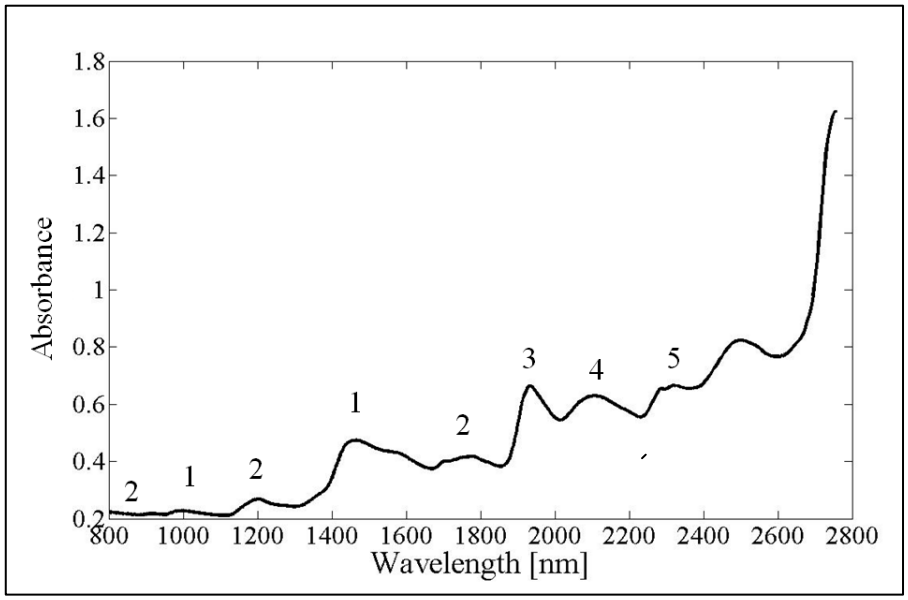

Figure 4.7: NIR reflectance spectrum of typical bread flour. Selected vibrational bands assigned to: (1) O–H and N–H stretch, (2) C–H stretch, (3) O–H combinations and N–H combinations, (4) amide and (5) C–H combinations**.**

Given that the interest variable analyzed was protein, regions pointed in NIR spectrum are expected to possess some correlation with the vibrational bands of functional groups presented in protein molecules. The match between the experimental results and theoretical knowledge is considered an indicator of this methodology's soundness.

Comparing results presented in Figure 4.6(b) with theoretical information in Figure 4.7, we conclude that NIR spectral regions indicated by ACO pheromone trail, as the most indicated for protein measurements, are associated with peaks 2 and 4 in Figure 4.7. This indicates a strong correlation between protein content and C–H stretch bands and the band associated with amide. We had expected the C-H stretch, as well as the region correspondent of peptide stretch, since both are functional groups presented in protein molecules and are considered key stretches for protein characterization (more descriptions about molecular characteristics of proteins can be found on Champe and Harvey (2005)).

The results obtained confirm the viability of applying pheromone concentration evolution during ACO in a PSCM routine as a useful tool for filtering and selecting spectral significant regions as well as discarding non-significant information in NIR spectroscopy.

PCR and PLS methodologies follow the same procedure previously applied, but using triplicate NIR measures as single measurements. In this case, the experimental data set is composed of 102 samples (34 flour samples measured in triplicate).

Considering that PCR and PLS are data selectors concentrating significant information from spectral data in just a few elements, the significant information is distributed in singular spectral regions. The selection of these specific spectral components allow the exclusion of possible noise sources and errors, beyond the increase in measurement velocity, robustness and data storage efficiency.

The similar results obtained from PCR and PLS prediction using full and filtrated NIR data set indicate that the methodology applied for the selection of spectral components is efficient and mainly selected the spectral elements that carry the most significant information correlated to the state variable.

Figure 4.8 presents the mean of pheromone trail obtained for protein characterization with NIR using PSCM and ACO. The horizontal lines represent a threshold, indicating the minimal pheromone concentration presented by the elements selected for 8.7% (line 1), 17.4% (line 2), and 43.5% (line 3) spectral components. The lines indicate the filtered data set obtained.

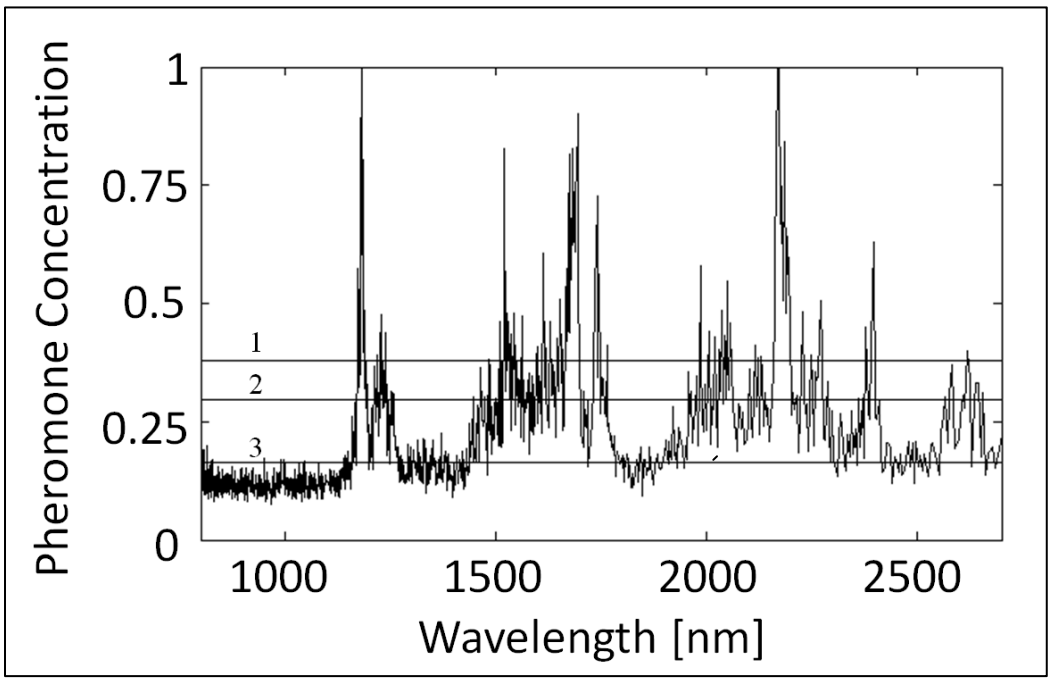

Figure 4.8: Mean of pheromone trail concentration (dimensionless) after search for groups composed of one to ten NIR spectral elements to predict protein in wet flour. The horizontal lines indicate the minimum pheromone concentration consider for filtered data: 8.7% (1), 17.4% (2) and 43.5% (3)**.**

The results presented in Figure 4.8 are directly correlated with the ones presented on Figure 4.6(b) (Figure 4.6(b) presents one of the ten groups of data used to generate the results presented in Figure 4.8). The similarities between those two figures highlight the convergence and tendency of ACO to attribute more importance (pheromone concentration) to specific spectral regions.

Comparing Figures 4.8 and 4.6(b), we can see that the use of mean pheromone trails highlight the main spectral regions instead of only few components, as occurs in Figure 4.6(b).

Using the importance ordering based on Figure 4.8, PCR and PLS models were generated and compared with full spectral models. Figures 4.9 and 4.10 show results obtained with all spectral data ranges, where only prediction results are presented. Figure 4.9 presents the results for PCR modeling while Figure 4.10 is related to PLS modeling.

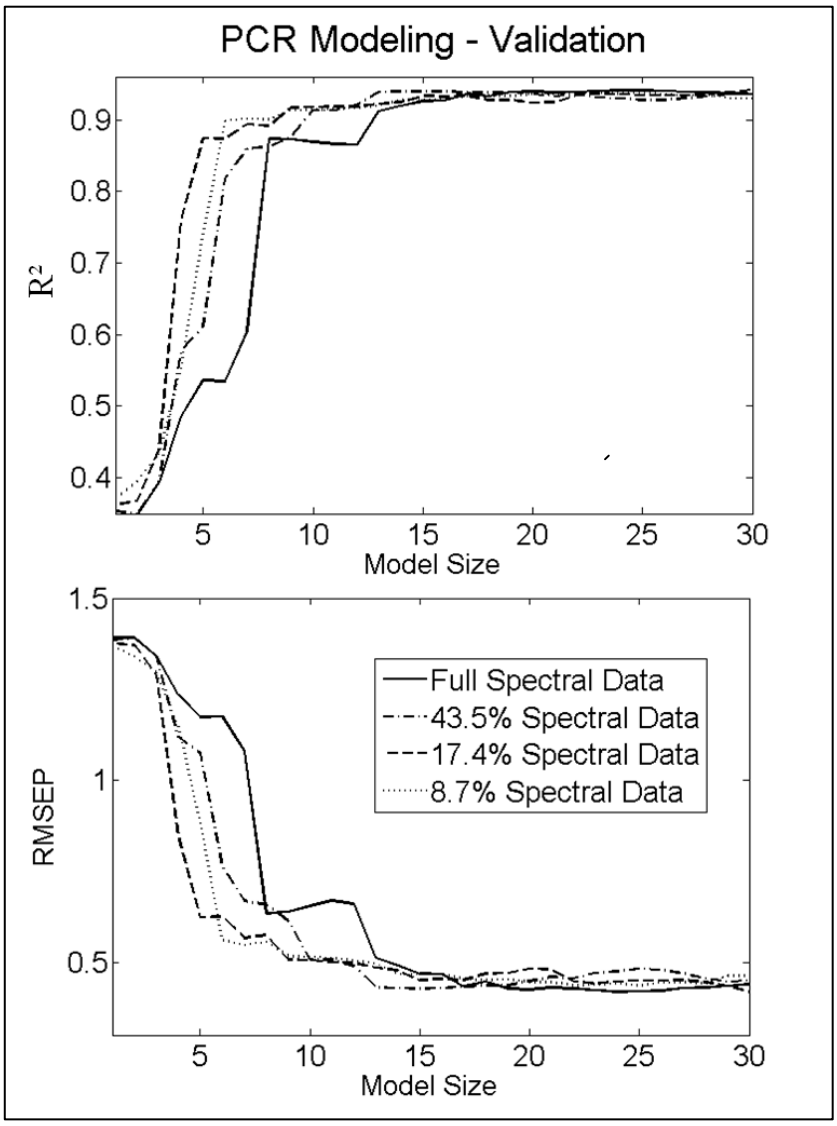

Figure 4.9: Coefficient of Determination (R^2) and RMSEP for validation phase of PCR modeling using full data set and filtered data set of NIR for protein prediction.

The results presented in Figures 4.9 and 4.10 shows that models of PCR and PLS obtained using filtered NIR data present more efficiency in condensing useful information for protein prediction in flour samples. This is confirmed by the fact that the sum or residual errors and determination coefficient achieve better values than the use of full spectral data. For instance, the result obtained with PCR modeling using filtered data works best for models smaller than fifteen input variables. For models with more than fifteen input variables, the result presented by filtered and full spectral models is equivalent.

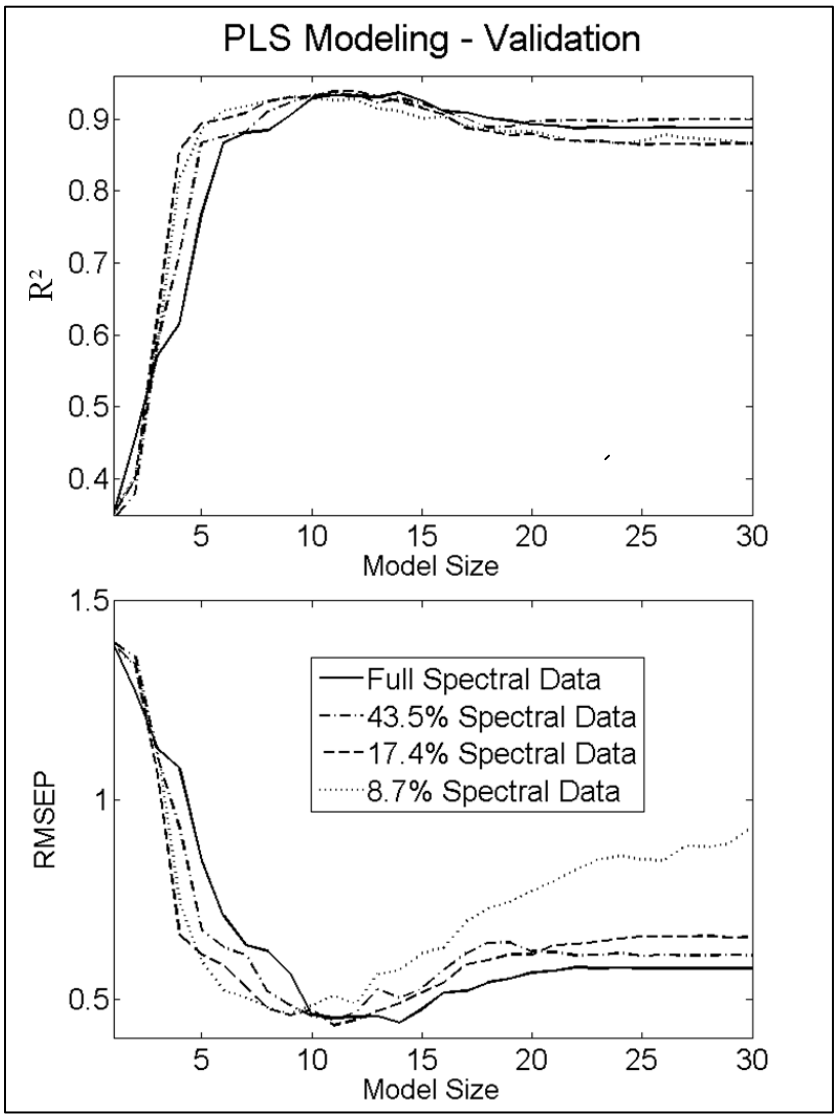

Figure 4.10: Coefficient of Determination (R^2) and RMSEP for validation phase of PLS modeling using full data set and filtered data set of NIR for protein prediction.

The results presented in Figures 4.9 and 4.10 are similar, proving the concentration capability of PSCM spectral data, given that prediction quality either increased or produced the same results for PLS and PCR models with the use of filtered data, especially for models with a smaller number of inputs.

However, PLS filtered models with more than ten input variables presented an increase in RMSEP, while standard models kept constant, although the values of R^2 remained equal to standard models. This indicates that even with the increase on RMSEP for PLS, the filtered methodology allowed the achievement of more accurate models using less number of input variables.

PCR and PLS models obtained using up to 17.4% of NIR data demonstrated better prediction capability than other models with less input variables, confirming the efficiency in selection of significant spectral variables and discard of noise or others factors that interfere in protein prediction.

Figure 4.11 presents the differences in model prediction obtained through filtering based on ACO pheromone trail using statistical coefficient of determination of RMSEP presented by full spectral modeling as standard references, evaluating the difference with the use of filtered data sets.

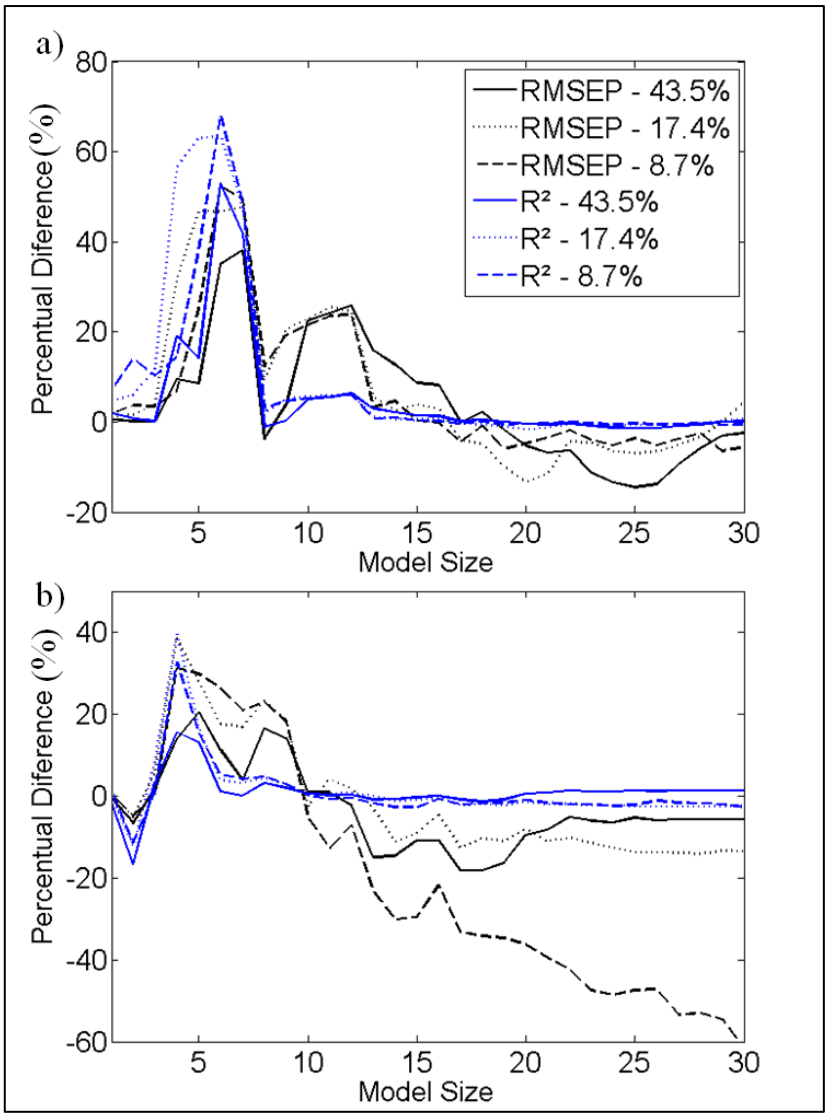

Figure 4.11: Percentage difference of RMSEP and R^2 between models obtained using full spectral data (reference value) and compressed data in PCR (a) and PLS (b) modeling.

The results presented in Figure 4.11 show that the use of filtered data sets improves significantly the prediction capability of PCR and PLS models for regression coefficient or RMSEP parameters. The increment in prediction precision is significant in models with up to ten input variables; after that, the models begin to show an increase in RMSEP, more pronounceable in PLS models, but nevertheless producing acceptably low absolute values.

The increase in prediction capability presented by PCR and PLS models provides evidence of ACO filtering capability in the selection of significant spectral components and rejection of spectral noise, usually responsible for the decrease in PLS and PCR variable prediction.

In relation to filtering rate, models that used 17.4% and 8.7% of full spectral data presented better results, although equivalent for models with up to six input variables. Models generated using 8.7% of spectral data presented better results than others models. This is a good indication that spectral noise rejection is effective and that we are only inserting significant information into PCR and PLS models.

4.4 Conclusions

The huge amount of spectral information provided by optical measurement techniques and the high sensitivity and correlation between spectrum components hinders our task of transforming information into process knowledge. Nonetheless, the application of multivariable modeling tools as PCR and PLS shows good results and applicability in this field.

When comparing PCR and PLS as tools for translation of NIR spectral measurements of flour samples in protein content information, both techniques achieve equivalent results, predicting the interest variable with the same degree of confidence. PLS, however, achieved better results using less independent variables compared to PCR. Therefore we conclude that PLS is more indicated to filter NIR data for protein content predictions in flour samples.

Furthermore, we obtain a more significant result when standard chemometrical methodologies are compared with PSCM. PSCM achieved higher precision in protein prediction than PCR and PLS. Models obtained using this methodology presented better results for any number of input variables, confirming that the selection and combination of pure spectral components, such as input variables on chemometrical models, presents some advantages when compared with methodologies that concentrate all spectral data into principal components or load vectors.

Among these advantages, we can mention: not needing complete spectral measurements, implicating viability of small on-line sensors based on spectral analysis and dispensing normalization of spectral data.

The importance of the spectral region map obtained using PSCM and ACO allowed spectral data filter as a function of interest variable. We applied pre-PCR/PLS data treatment to discard non-significant spectral regions, improving the chemometrical models precision once models are not affected by irrelevant information.

PCR and PLS are standard chemometrical methodologies presenting good results on variables prediction. However, the advantages presented by PSCM in chemometrical modeling and spectral analysis shows how viable it is for on-line sensor development, not only to characterize individual samples but for process variable measurements as well as its usefulness as an analytical source of knowledge on processes.

The filtering rate of spectral data tested (reducing total amount of spectral components in 80%, 70% and 50%) significantly improved prediction of PCR and PLS models. For some model sizes, the improvement in efficiency surpassed 60% in relation to the same models structures that used full spectral data. Those improvements in results are due to the selection of right spectral region correlated with interest variable, and the

elimination of noise and secondary changes to spectrum caused by secondary state variables. Spectral data filtering provides a more precise representation that is bound to the desired state variable.

Based on the results of pre-filtered NIR spectra chemometric modeling using pheromone trail concentration evolution in Ant Colony Optimization procedures for protein prediction, the increase of robustness and prediction capability proved that the tendency to increment pheromone concentration around spectral components presents greater correlation with the state variable evaluated, allowing this information to be used as a pre-phase of spectral data treatment.

Capítulo 5 – Characterization of *Saccharomyces cerevisiae* fermentation using Fluorescence Spectroscopy 2D

Abstract:[*] For process control it is necessary to measure or estimate the controlled variable. This work proposes a new methodology for the characterization of batch fermentations of glucose with *Saccharomyces cerevisiae*, using Fluorescence Spectroscopy. Chemometric models were created and tested using real process data. We applied two different approaches in the modeling: (i) standard chemometrical methodologies (PCR and PLS); and, (2) Pure Fluorescence Spectral Components (PSCM). PSCM led to a more precise prediction of process state variables for most of the input variables analyzed. Additionally, PSCM methodology associated with Ant Colony Optimization (ACO) allows the selection of the most significant spectral regions for each one of the state variables. For instance, instead of using all spectra data (as when PCR and PLS is applied), PSCM identified the most important region between 270nm to 390nm for excitation and 310nm to 430nm for emission wavelengths. Applying the reduced fluorescence region increased the accuracy of standard PCR and PLS models prediction when compared with full spectral predictions. The batch fermentation illustrates the applicability of ACO and PSCM as a tool for characterization of Fluorescence Spectroscopy.

[*] *Submitted to Applied Spectroscopy*

5.1 Introduction

Biotechnological processes are taking on a prominent position in the production matrix and have grown proportionately more than the standard chemical process (Mussatto, Dragone *et al.* 2010). An increase in production, analogous to chemical processes, can be achieved by applying control strategies designed to improve the process operation. Success, however, is closely related to knowledge of process states, and, because of the lack of concentration sensors, this represents a serious problem for bioprocesses (Ranzan, Trierweiler *et al.* 2011).

Historically, the most effective way to improve production in a bioprocess plant has been associated with the evolution of strains used in fermentation processes (Aynsley, Hofland *et al.* 1993). However, recent developments show that improvements in the bioprocess can be obtained by using supervision and control tools, reducing production costs while maintaining the quality of the desired products (Yamuna and Ramachandra 1999, Whitford and Julien 2007, Kabbaj, Nakkabi *et al*. 2010, Menezes 2011).

Optical sensors are state-of-the-art technology in the characterization of bioprocesses. These sensors present the most interesting characteristics for application on the bioprocess, such as reliability, precision, low dead time, high sensitivity, with the further advantage that they do not need to remain in contact with the reactional media, drastically reducing contamination risks (Whitford and Julien 2007).

Inside the universe of optical phenomenae, various analytical methodologies can be applied to development of sensors, for instance, Near-infrared spectroscopy (NIR), Raman Spectroscopy, Fluorescence Spectroscopy, among others (Whitford and Julien 2007). Specifically, 2D Fluorescence Spectroscopy is considered one of the most promising methodologies for bio-sensor development, due to its high sensitivity to metabolites, fast measurement and huge amount of known fluorophores (Horvath, Glazier *et al.* 1993, Stärk, Hitzmann *et al.* 2002).

Fluorescence Spectroscopy allows the monitoring of intracellular metabolites (e.g.: NADH, amino acids, proteins, vitamins and others) in a non-invasive way, through the combination of wavelengths (excitation/emission fluorescence pairs), to measure 2D fluorescence spectrum (Stärk, Hitzmann *et al.* 2002).

The critical problem associated with the conversion of spectral data into state variables lies on the huge amount of information contained inside the spectra. This problem requires the application of virtual analyzers or state estimators to translate the fluorescence information into process variable knowledge in order to determine the relationship between fluorescence data and process variables (Boehl, Solle *et al.* 2003). Among the chemometric modeling techniques, applied in data processing, Partial Least Squares (PLS), Principal Component Regression (PCR) and Neural Networks are standard methodologies (Christensen, Norgaard *et al.* 1995, Wolf, Almeida *et al.* 2001, Hagerdon, Legge *et al.* 2003, Solle, Geissler *et al.* 2003).

PCR is not directly applied in the original spectral data, but is correlated to the PCs (Principal Components) obtained by PCA (Principal Component Analysis). Once the PCs are mutually orthogonal, the typical problem of collinearity and high correlation, which arises in many regression techniques, can be avoided. In this methodology, the PCs are

combined in a way to predict the output data matrix, using linear multivariable regression (Liu, Kuang *et al.* 2003).

PLS is a multivariate statistical technique with the goal of correlating two data sets and predicting one set based on another. It attempts to identify not only the factors (latent variables, LVs) that capture the largest amount of variance in data matrix, but also those that allow a linear correlation to be obtained between the spectral data and the process variables (Geladi and Kowalski 1986).

This paper focuses on evaluating the viability of 2D fluorescence spectroscopy as a tool for on-line monitoring of glucose fermentation with *Saccharomyces cerevisiae*, comparing the standard chemometric methodologies with a self-developed method based on Pure Spectral Components Modeling (PSCM). In this case, we only use pre-selected spectral components for state variable prediction, simplifying the spectral analysis and reducing the noise and error propagation when full fluorescent spectral measurements are used for prediction. In this area, the main contribution of this work is the mapping of 2D fluorescent spectroscopy data in function of fermentative significant variables, allowing pre-treatment of fluorescent data for variable inference through customized sensors.

5.2 Materials and Methods[1]

5.2.1 2D Fluorescence Spectroscopy

Fluorescence Spectroscopy, also known as Spectrofluorometry, is a technique that analyses the fluorescence of fluorophores contained in a sample, as a result of the emission/re-emission of low energy light, usually in ultraviolet wavelengths. It is based on the re-emission of light with spectral change, which some chemical species present. The re-emitted light is proportional to the fluorophore concentration in the analyzed medium and has a wavelength equal or superior to the excitation (Hitzmann, Pekeler *et al.* 1998, Solle, Geissler *et al.* 2003).

In practical terms, in fluorescence spectroscopy a continuous source of light shines onto an excitation monochromator, who selects a wavelength band. The monochromatic excitation light is directed onto a sample, which emits luminescence. This luminescence is directed onto a second emission monochromator, which selects a wavelength band and shines it onto a detector. The equipment is generally connected by fiber-optic bundle to a glass window in a reactor, making this a non-intrusive, non-invasive and real time measurement method (Scheper, Hitzmann *et al.* 1999, Hantelmann, Kollecker *et al.* 2006).

The fluorescence experimental data used in this work consists of two cultivations of glucose by *Saccharomyces cerevisiae* H620 growing in a 1.5L bioreactor at constant temperature and pH, 30ºC and 5.5, respectively, with Schatzmann medium supplementation. During cultivation, we collected fluorescence spectra every 6 minutes, using a BioView fluorometer (Delta Light & Optics, Denmark), as described by Stärk *et al.* (2002). Each spectrum contained 150 fluorescence pairs with excitation/emission

[1] A descrição detalhada deste item encontra-se no Capítulo 3

wavelengths: 15 filters in the region of 270 to 550 nm for excitation and 15 filters in the region of 310 to 590 nm for emission, both with a bandwidth of 20 nm, collected equidistantly.

We collected 190 spectra from each cultivation. The data obtained by BioView Spectrum Fluorometer was processed with MATLAB software (Ver. 5.3.0.10183 R11, The Mathworks, Inc., Natick, USA).

5.2.2 Dynamic Modeling of Saccharomyces cerevisiae Fermentation

In order to effectively evaluate chemometric models based on fluorescence data, information about state variables must be available at the same sample range of fluorescence data. Since, originally, the total amount of off-line data is considerably smaller than fluorescent sampling data, we simulated a dynamic model of the fermentative system to interpolate the state variables, obtaining off-line data at the same sampling range as fluorescent data.

When *S. cerevisiae* cells are exposed to glucose in a medium propitious for their growth, they produce ethanol, carbon dioxide and biomass, by which one can observe a diauxic growth pattern. This phenomenon is characterized by cellular growth in two phases during the batch fermentations (Zang, Scharer *et al.* 1997).

The first growth phase is characterized by consumption of glucose by fermentative pathway, producing ethanol. When glucose is no longer available, *S. cerevisiae* cells begin to consume ethanol in the aerobic pathway (second phase) (Zhang, Zamamiri *et al.* 2002).

Mathematically, diauxic growth is considered a discontinuity in the dynamic model. This phenomenon denotes a change of the primary substrate supply (from glucose to ethanol) used by the cells. In the presence of glucose, yeast generally recognizes the carbohydrate as the substrate, but when the medium becomes low in glucose, the biomass modifies its metabolism to use ethanol as substrate.

When this process is operated in batch mode, it can be modeled by:

$$\boldsymbol{dG}/\boldsymbol{dt} = -\boldsymbol{X}.\boldsymbol{\mu_G}/\boldsymbol{Y_{GX}} \tag{5.1}$$

$$\boldsymbol{dE}/\boldsymbol{dt} = \boldsymbol{X}.\boldsymbol{\mu_G}/\boldsymbol{Y_{GE}} - \boldsymbol{X}.\boldsymbol{\mu_E}/\boldsymbol{Y_{EX}} \tag{5.2}$$

$$\boldsymbol{dX}/\boldsymbol{dt} = \boldsymbol{X}.\boldsymbol{\mu_E} + \boldsymbol{X}.\boldsymbol{\mu_G} \tag{5.3}$$

Whereby we disregarded the production of carbon dioxide and considered concentrations of glucose (*G*), ethanol (*E*) and biomass (*X*). μ_G and μ_E as the specific growth rates in glucose and ethanol, respectively. Y_{GX}, Y_{GE} and Y_{EX} are the specific yields of: biomass from glucose, ethanol from glucose, and biomass from ethanol.

The growth rate values (μ_G and μ_E) are considered a function of glucose concentration ($\mu_G(G)$ e $\mu E(_G)$). To this end, one could use expressions commonly found in the literature to describe this dependence (Monod, Aiba, Levenspiel, amongst others) (Mulchandani and Luong 1989, Habibi, Vahabzadeh *et al.* 2013); however, in this study, we set forth a new form of dependence between growth rates and glucose concentration, where those variables are correlated by a hyperbolic tangent factor, as described by eqs. (5.4) and (5.5). Growth rates are dependent on glucose concentration and four constant parameters (a, b, μ_{Gm} and μ_{Em}).

$$\boldsymbol{\mu_G = \mu_{Gm}\left(0,5 - 0,5tanh\left(a(b - G)\right)\right)} \quad (5.4)$$

$$\boldsymbol{\mu_E = \mu_{Em}\left(0,5 - 0,5tanh\left(a(G - b)\right)\right)} \quad (5.5)$$

In a dynamic model, diauxic growth results in a change between μ_G and μ_E. When glucose is available, μ_G values are greater than zero. On the other hand, when all the glucose is consumed and the biomass changes its metabolism, μ_G value is equal to zero and μ_E (which has a value equal to zero when glucose is available) has values greater than zero.

In order to solve the dynamic system formed by eqs. (5.1) to (5.3) we applied the Runge-Kutta method. In all simulations, the coefficients of specific yield are the ones reported by Solle *et al.* (2003) (Y_{GX} = 0,167 g_{cell}/g_{gluc}, Y_{GE} = 0,5g_{etha}/g_{gluc}, Y_{EX} = 0,333 g_{cell}/g_{etha}). We estimated a, b, μ_{Gm} and μ_{Em} using SIMPLEX methodology and the experimental data of both fermentations resulting in μ_{Gm} = 0.3792 h^{-1}, μ_{Em} = 0.0587 h^{-1}, a = 1.3968 $L.g^{-1}$ and b = 2.2136 $g.L^{-1}$. Figure 5.1 presents the original and the respective simulated off-line data for the fermentations.

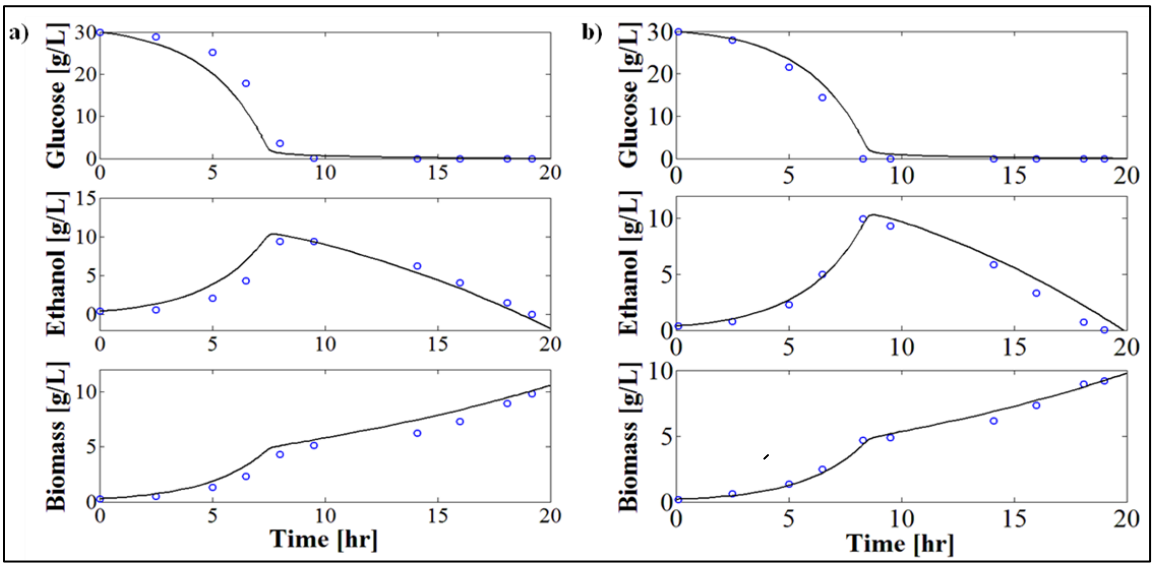

Figure 5.1: Fermentation 1 (a) and Fermentation 2 (b). (o) Off-line data. (—) Simulated data using dynamic model.

5.2.3 Pure Spectra Chemometric Modeling (PSCM)

Based on the pre-selection of spectrum elements (e.g. wavelengths in NIR and Raman Spectroscopy or excitation/emission wavelengths pairs in 2D fluorescence spectroscopy), the proposed method aims to choose spectrum components that, when analyzed together, present a high correlation with state variables, discarding spectral noise and errors added by spectra regions not significantly related to the interest variables (Skoog, Holler *et al.* 2007).

Many strategies can be applied in the selection of spectral components (e.g. independent correlation between spectral components and state variable, variance of spectral components, etc). In this work, the selection of spectral groups is made using Ant Colony Optimization (ACO), where the best combination of spectral components is sought based on the prediction error of a calibration data set with known values of interest variables and associated spectra. This strategy develops the simultaneous mapping of the significance nexus of spectral data and state variables, a feature that can be applied for highlighting significant spectral regions, allowing the theoretical and practical understanding of active fluorophores and fluorescence mechanics of the process system.

The version of ACO implemented in this study is based on pheromone trail evolution during spectral group scanning. Initially, all spectral components are marked with the same pheromone concentration (Ranzan, Strohm *et al.* 2014). The ACO routine selects random spectral components for a compound test group that is evaluated using the objective function for process variable prediction. Based on objective function error, the pheromone concentration, associated with each spectral component at the evaluated spectral group, is updated. For the subsequent spectral group selection, the random selection chooses spectral components associating the same random trigger and a cumulative density of pheromone for the full range of spectral elements.

The combined association of random selection and pheromone density brings into evidence significant elements inside the spectral range, and, after few iterative runs, a pheromone profile is established, and a density pheromone trail highlights the significant spectral for process variable prediction. See (Ranzan, Strohm *et al.* 2014) for more detail about this algorithm.

5.2.4 Chemometric Modeling Procedure

In order to test the fermentative process characterization using 2D fluorescence spectroscopy, the chemometric protocol used is based on chemometric models calibration and models test. To evaluate a reliable analysis of prediction capability and robustness, adjusted models are calibrated and tested with distinct sets of experimental data.

Given that the amount of experimental data available englobes two batch fermentations, we segmented experimental data so that first fermentative batch is used only in the models calibration phase, and fermentative batch two is used only in the model test phase. This way, we can guarantee that the calibration and test of chemometric models are performed with distinct experimental data and results are directly associated with robustness of predictive models in real process application.

5.3 Results and Discussion

5.3.1 Qualitative Data Analyses

First, we applied PCA analysis to verify the similarity of spectroscopy fluorescence data obtained from two cultivations. Figure 5.2 shows the first two principal components, and indicates that both spectral data are similar and no pre-processing is needed to perform a comparison between fermentations. In other words, these fermentations have equivalent spectrum ranges, indicating the same characteristics regarding medium composition and process operation. Variations on those variables could reflect significant variance at spectrum characteristics, impacting models' prediction capability. Thus, spectrum data should be mathematically treated so that they are equivalent.

Given that the efficiency of PCR and PLS methodologies are highly associated with spectral data quality, it is useful to normalize the spectral signals prior to data analysis. This process helps in eliminating arbitrary offsets and multiplication factors. We achieved this by applying Standard Normal Variate (SNV) scaling to spectral data. This method essentially autoscales the samples, obtaining zero mean and standard deviation of 1 for each spectrum (Gemperline 2006, Wehrens 2011).

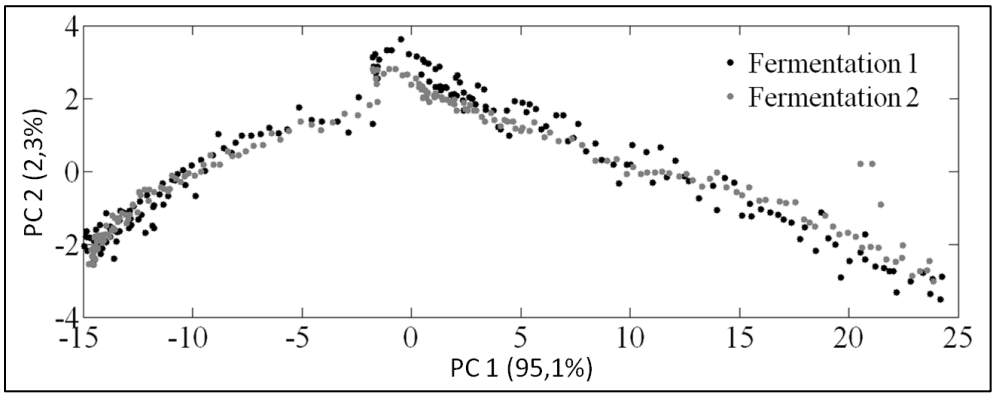

Figure 5.2: First Principal Component versus Second Principal Component of fermentations 1 and 2.

5.3.2 Characterization of Glucose fermentation using Fluorescence Spectroscopy

The chemometric modeling for state variable prediction is made using the previously described methodologies - PCR and PLS - as standard methodologies and PSCM for improving prediction. We apply these methodologies to SNV normalized fluorescence spectral data. As described before, we obtained off-line data from simulation.

For comparing chemometric model predictions, we performed the following steps:

1) Normalization of spectral data;

2) Model calibration (using simulated fermentation 1 off-line data at same sampling rate as 2D fluorescence spectroscopy);

3) Validation Test (using simulated and spectroscopic data from fermentation 2).

Step 3 provides information about the applicability of fluorescence for characterization of this particular fermentative process.

Once the structures of PCR, PLS and PSCM models are polynomials multilinear at parameters, fitting is made through Ordinary Least Square (OLS) problem. This problem has an algebraic solution and model parameters are calculated with equation 5.6, where *X* is the independent variables matrix, *Y* is the dependent variable vector and *β* is the vector of model parameters.

$$\boldsymbol{\beta} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Y} \tag{5.6}$$

Two statistical metrics are used as performance indices, RMSEP (Root-Mean-Square Error of Prediction) and R^2 (Coefficient of Determination), calculated according to equations 5.7 and 5.8, respectively. In equations 5.7 and 5.8, the sub index *p* refers to the vector of prediction, *m* to the measured variable vector and *N* is the number of measurements.

$$\boldsymbol{RMSEP} = \sqrt{\sum_{i=1}^{N}\left(y_{p,i} - y_{m,i}\right)^2 \Big/ N} \tag{5.7}$$

$$\boldsymbol{R^2} = \mathbf{1} - \sum_{i=1}^{N}\left(y_{p,i} - y_{m,i}\right)^2 \Big/ \sum_{i=1}^{N}\left(y_{m,i} - \overline{y}_m\right)^2 \tag{5.8}$$

Figure 5.3 shows the statistical indices of chemometric models at step 3. This figure presents the results of models calibrated using fermentation 1 data and tested on fermentation 2 data, grouped by indices (R^2 and RMSEP) and state variables (Glucose, Ethanol and Biomass concentration). The factor selected for comparing between models was the number of input variables. Models obtained with PCR can be compared with models obtained by PLS or PSCM, which use the same number of independent variables (PCR – Principal components, PLS – Load Vectors, PSCM – Pairs of Fluorescence), and, consequently, the same number of estimated model parameters, for the prediction of the dependent variable, given that the model structures applied in all methodologies are similar (multi-linear models).

For the process data analyzed, all tested chemometric methodologies achieved accurate results in the prediction of a set of simulated interest variables for fermentation 2 using spectral fluorescence data from this same fermentation. The fact that all chemometric methodologies resulted in a satisfactory proximity to process data indicates that 2D fluorescence spectroscopy is a reliable way to monitor state variables on glucose fermentation.

2D Fluorescence spectra carry a significant amount of information about process, which leads to the problem of how to select information inside the spectra and of selecting the more appropriate method of translating spectroscopic information into state variables knowledge.

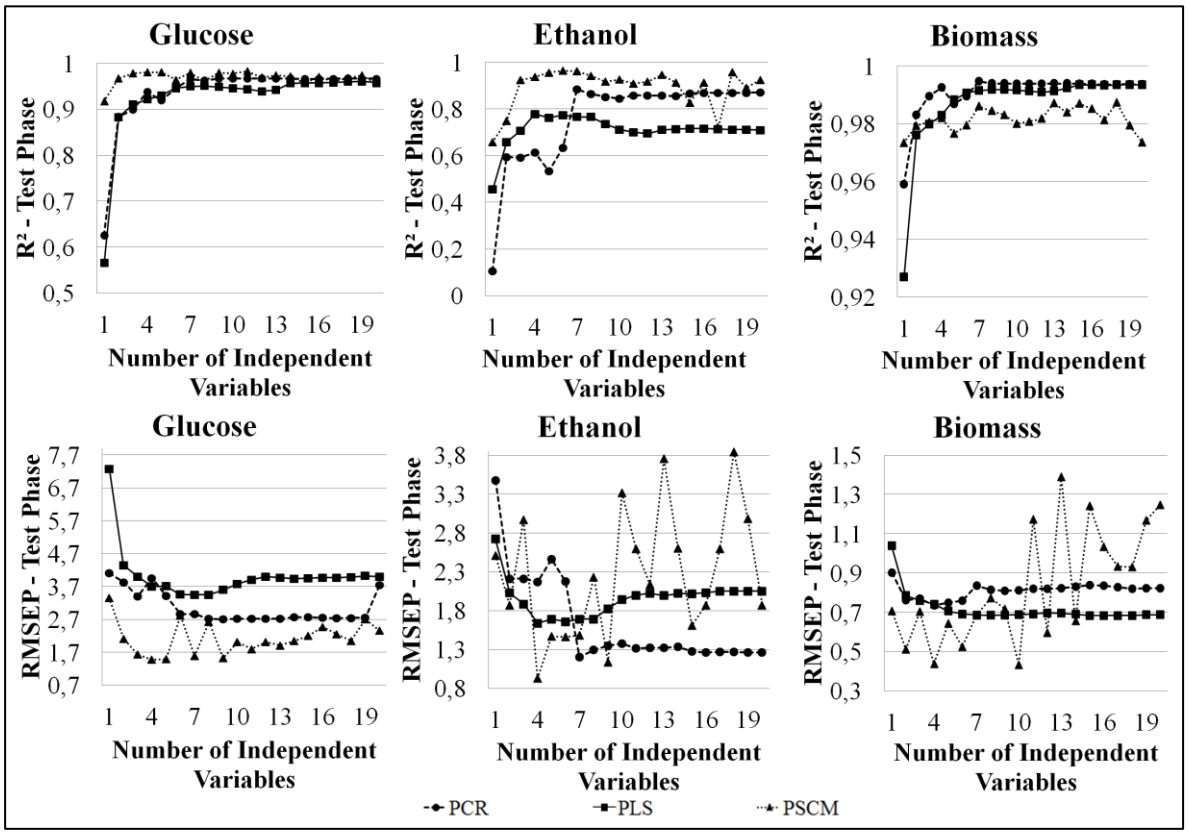

Figure 5.3: RMSEP and R^2 versus number of Input variables in PCR (principal components), PLS (load vectors) and PSCM (pairs of fluorescence) models for Ethanol, Glucose and Biomass prediction of fermentation 2, using data from fermentation 1 for model calibration**.**

All chemometric methodologies presented similar results in the prediction of Glucose and Ethanol content in fermentation 2, especially those with more than five independent variables, as R^2, achieved equivalent values. For models with less than five input variables, the results obtained with PSCM models were significantly better than the ones achieved using PCR and PLS.

Biomass prediction was more accurate using standard methodologies. However, all methodologies were able to accurately predict biomass content, achieving R^2 higher than 0.97, and RMSEP less than 1.5, for all models with more than one input variable.

PSCM prediction shows variable results of statistical indicators, as a function of model size. This behavior is probably due to fact that ACO is classified as a heuristic optimization method, meaning that the number of spectral combinations increases significantly with the increase in model size, raising the possibility that optimization algorithms stop at objective local minima. One should determine a greater number of iterations for

optimization to overcome this issue. Another possible explanation for that behavior is due to overfitting models parameters to calibration data group. Despite validation oscillatory behavior, results are satisfactory for all interest variables prediction, confirming PSCM modeling using fluorescent data, as fermentative characterization option.

Results suggest that any of the tested methodologies can be applied to predict Glucose or Biomass concentration; however, considering the size of models and quality of prediction, the results obtained using PSCM were superior in indicating Glucose and Ethanol inference models.

This superior result of PSCM was aligned with our expectation, given that only information directly associated with the studied variables are selected, whereas, on PCR and PLS, the principal components and load vectors are loaded with all significant information on fluorescent spectrums, making PSCM models specific to interest variables.

Regarding ethanol prediction, the PSCM models presented the best results, with the exception of the models with 14 and 16 parameters (regarding R^2 indices). However, the values of RMSEP for PSCM presented worse results than PCR and PLS, indicating that the standard error presented by PSCM models was larger compared with the other methodologies. Nonetheless, this difference is not enough to impact R^2, and it is still considered the best methodology for ethanol characterization.

Since PSCM presented good or, in some cases, better results than standard chemometric methods, sensors based on small amount of fluorescent pairs appear to be viable and more reliable than sensors based on full range spectrum. This is a significant result for biochemical and chemical characterization allowing the development of customized sensors that can help to improve process monitoring and control.

5.3.3 Spectral data Characterization

During PSCM modeling, ACO is applied as a selection tool for fluorescent pairs to process variables prediction, generating an internal variable called “Pheromone Trail”. It contains the amount of importance ascribed by ACO to each spectral component in the prediction of the interest variable.

The pheromone trail is a vector associated with all spectrum components. Initially, it has the same value for all components. Then, the values change during the optimization routine in a discrete manner (element by element), in such a way that elements with greater accuracy in interest variable prediction receive a greater pheromone amount. In this way, pheromone trail acts as a qualitative filter based on trial and error during optimization runs. At the end of the routine, global or local minima are achieved and pheromone trail presents its signature associated with interest variable, emphasizing significant data components.

This dimensionless variable has the advantage of providing reliable information to characterize the spectral data based on state variables, allowing the segmentation of fluorescence spectrum into regions more related to the state variables.

Fluorescence measurements are well known as a highly sensitive analytical technique, able to distinguish between small variations of medium concentration. If the applied

chemometric methodology is not efficiently treating spectral noise created from medium components unrelated with the interest variable, problems may arise. PCR and PLS use the variance of principal components or load vectors, respectively, as indicators to translate spectral data and predict state variables. They are highly dependent of spectral data pre-treatment to block and filter spectral noise, but are unable to distinguish between state variable of medium noises responsible for spectral changes.

The pheromone trail can be explored as a filter, condensing the fluorescent data into a subset that contains only the significant information for the interest process variable. This filtering process reduces the amount of non-significant information within the model, reducing the influence of noise in the models.

Figure 5.4 shows the pheromone trail information, normalized between one and zero for excitation range from 270nm to 430nm, and emission range of 10nm to 430nm. The highlighted regions on each graphic are associated with the most significant fluorescent pairs assignment with their respective state variable.

In this study, we propose to use the ACO pheromone trail as a tool for search and selection of the spectral regions that are directly correlated with each state variable, enabling the characterization of process variables based on spectral changes only in specific regions. This features acts like a searching engine for state variable signatures inside the fluorescence spectra.

Figure 5.4 indicates that characterization of all interest variables from glucose fermentation can be done through the analysis of a relatively small region of fluorescence spectrum. This is a valid property from the sensor development point of view, because the spectral range of light sources and detectors is reduced. On the other hand, it can lead to a model development issues, because the models could lose precision due to fluorescence pairs being simultaneously used for distinct state variables prediction and these variations can mask fluorescence changes on the others variables.

The problem of variables' cross interference on model prediction is not specific to this fermentative process, since PCR, PLS and PSCM models presented acceptable results in the prediction of all tested fermentative variables.

As shown in Figure 5.4, glucose presents a larger significant region (usually associated with protein fluorescence measurements) than ethanol (region associated with NADH fluorescence) and biomass (regions associated with tyrosine fluorescence). Furthermore, this region comprehends parts of the important spectral regions correlated with ethanol and biomass. Nonetheless, both variables also present other specific regions that are important for its prediction as well. Therefore, by associating the equivalent region with the specific region for each variable, one could obtain distinct models for each variable.

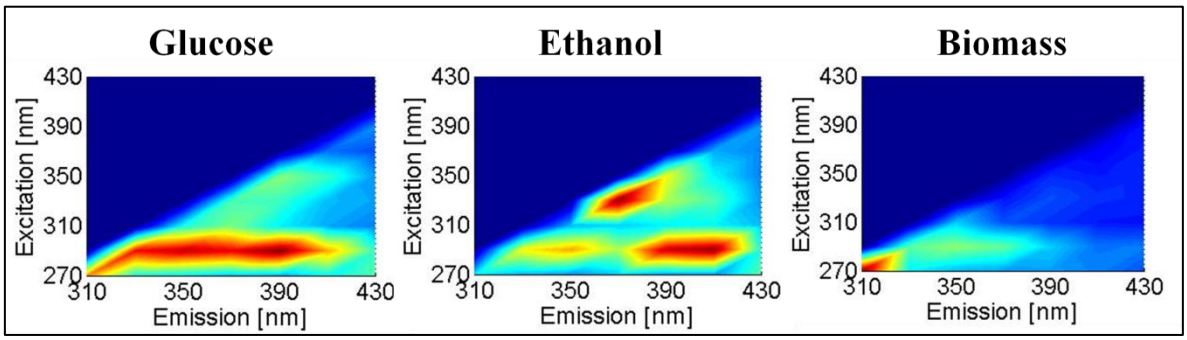

Figure 5.4: Significant fluorescence spectral regions associated with Glucose, Ethanol and Biomass concentration, obtained by ACO pheromone trail evolution.

The efficiency of ACO spectrum regions classification is tested using the region between 270nm to 390nm of excitation and 310nm to 430nm of emission (region containing most of the significant information for all state variables) for calibration and testing PCR and PLS chemometric models. Once PCR and PLS condenses all the significant information in principal components and load vectors, the comparison between models based on full range spectral data and reduced spectral data could provide evidence that the selected region contains the main information for state variable characterization inside full fluorescence spectra.

The reduced spectral region applied in modeling correspond to approximately 16.5% of the full fluorescence spectral data previously used on PCR and PLS models. In this way, the capability of new chemometric modeling in predicting satisfactory state variables is directly correlated with the selection of the correct spectral region that contain the right information about the desired state variable.

Results presented on Figure 5.5 prove that the primary information associated with each state variable is effectively found inside the region selected by ACO. The results of R^2 for the test phase of PCR and PLS modeling have greater, or at least similar prediction success, when reduced spectral data are compared to full spectral data models. This result confirms the applicability of ACO for filtering and selecting significant spectrum selection prior to applying PCR and PLS.

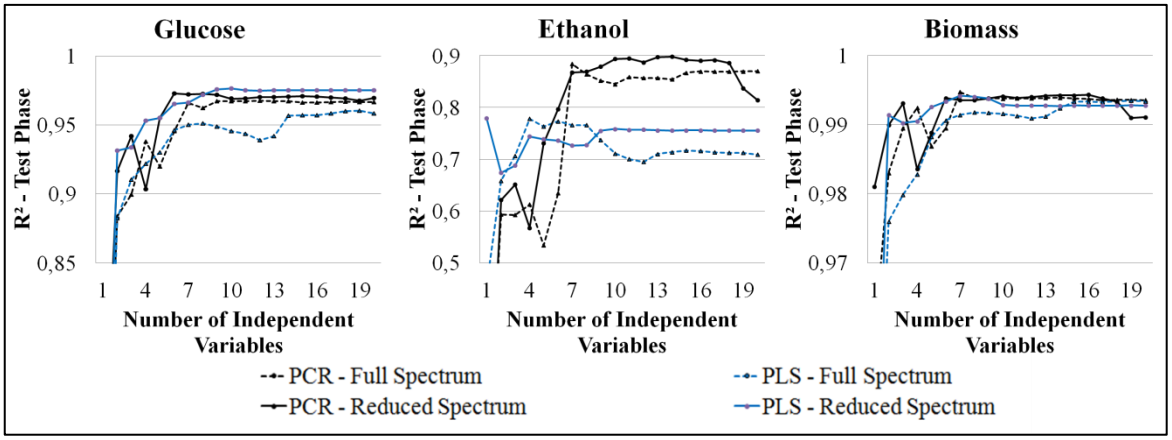

Figure 5.5: PCR and PLS modeling using Full Fluorescent Spectral data and Reduced Fluorescent Spectral data based on ACO analysis.

This feature presented by ACO is useful for theoretical analysis of fluorescence spectra, especially when they are obtained from complex mixtures. The spectrum segmentation into regions allows the development of specific chemometric models for each compound, aiming to minimize the spectral influences of other compounds in the model.

5.4 Conclusions

In fermentation processes with *Saccharomyces cerevisiae* operating in batch configuration, simple dynamic models are usually applied for simulation, although these models cannot always represent modifications in process dynamics caused by metabolism changes.

The dynamic simulation of off-line data performed in this study enabled the efficient modeling of diauxic growth. The modified dynamic model included terms in the growth rate equations dependent on glucose concentration, taking into account changes in metabolism when glucose is completely consumed.

The simulation of off-line variables is an important tool that allows the evaluation of fluorescence spectral data for the development of reliable on-line chemometric models for state variable predictions, providing off-line measurements (by simulation) at the same sampling rate as the spectral data.

The results of PCR and PLS chemometric models for predicting Glucose, Ethanol and Biomass content using 2D Fluorescence Spectroscopy confirmed that this analytical method can be applied in the development of inference sensors for on-line fermentation characterization. The results of models during the test phase presented high values of R^2 and low values of RMSEP.

Analysis of spectral data using PSCM has shown a better result for chemometric modeling when compared with PCR and PLS models, showing that the use of pure spectral signals to prediction of state variables is more accurate. The better results presented by PSCM are due to the fact that when groups of spectral components are selected, cross interference caused by noisy regions or secondary process variables are neglected.

Unlike PCR and PLS, PSCM models do not require pre-treatments of spectral data (e.g. normalization or standardization). This feature of PSCM models, combined with the fact that a small amount of spectral components presented good results in variable predictions, shows that development of small sensors based on just a few fluorescence wavelength measurements are viable and can be directly applied on a real process, possibly in parallel with standard methodologies already applied within the industry.

From a theoretical and practical point of view, results provided by PSCM methodology associated with ACO (segmentation of spectral data into significant regions as function of state variables) is relevant not only for the development of specialized sensors, but for further developments of research in chemical and biochemical processes as well.

Knowledge on spectral regions associated with variables, metabolic behaviors, structural changes and other process parameters are a significant contribution for process instrumentation and characterization, features that are extremely necessary to the optimization and control of biotechnology processes.

Capítulo 6 – NIR pre-selection data using modified Changeable Size Moving Window Partial Least Squares and Pure Spectral Chemometrical Modeling with Ant Colony Optimization for wheat flour characterization

Abstract:* *The aim of process optimization is obtaining higher productivity and profit in chemical or bio-chemical process. For that, one must apply control techniques that closely correlate with our ability to characterize a process. Optical sensors associated with chemometric modeling are considered a natural choice for non-intrusive and high sensitivity measurements. This study focus on wheat flour characterization (usual and mandatory action, widely present on the food industry) using Near-Infrared, comparing two approaches for spectral region selection: modified CSMWPLS and PSCM/ACO. Spectroscopic data is assayed using a combination of CSMWPLS and variable selection algorithm based on Ant Colony Optimization. Protein prediction results are compared with standards PLS, CSMWPLS and PSCM/ACO models. Prediction capability improved 46% using modified CSMWPLS and PSCM/ACO modeling, confirming the efficiency of the proposed characterization methods and chemometric modeling strategy.*

*For submission

6.1 Introduction

Industrial needs for online monitoring and control of process key variables encourage the research and development of new methods for measurement (Whitford and Julien 2007). The limitations associated with sensors, regarding their expensive costs, or even its not reliable quality measurements, leads to the development of data-driven soft sensors, like those based on canonical variate analysis (CVA), partial least-squares (PLS), artificial neural networks (ANN) neuro-fuzzy systems and Gaussian process regression (GPR) (Ni, Brown *et al.*).

Nowadays many soft sensors are widely accepted and applied as viable and useful methods for online qualitative predictions of processes variables. Despite the huge available options, PLS regression is the one most adopted due its advantages for noisy and correlated data, usually common in industrial processes (Du, Liang *et al.* 2004, Brown 2013, Cariou, Verdun *et al.* 2014, Chi, Fei *et al.* 2014).

Partial least squares regression is a worldwide popular multivariate calibration method. It is applied in multi-component spectral analysis, fact that made it the first choice for studies and development of sensors based on vibrational spectroscopy, like infrared (IR), near Infrared (NIR), Raman Spectroscopy, and fluorescence (Hasegawa 2001, Du, Liang *et al.* 2004).

Despite the well know PLS capacity of dealing with full-spectrum calibration problem, the selection or filtering of spectrum regions is still a very important issue, once its impact on models prediction capability are directly associated with the sensitivity of spectral data to process medium changes and their influences in specific spectral regions (Xu and Schechter 1996, Sato, Kiguchi *et al.* 2004, Sratthaphut and Ruangwises 2012).

Jiang *et al.* (2002) made an extended and detailed discussion about the methods for spectral interval selection, proposing a method called moving window partial least squares regression (MWPLSR). That method searches for informative spectral regions for multi-component spectral analysis, using a series of PLS models, for spectral mapping, that prospects all spectral range using a continuous size moving window. A new model is obtained for each window displacement and tested using Cross-validation strategy, optimizing the spectral region for PLS interest variable inference.

Based on Jiang's *et al.* (2002), Du *et al.*, (Du, Liang *et al.* 2004) proposed an evolution of MWPLSR, introducing two new methods for spectral selection, the changeable size moving window partial least squares (CSMWPLS) and the searching combination moving window partial least squares (SCMWPLS).

Both methods proposed by Du *et al.*, (2004) are combinations of WMPLSR strategy. On CSMWPLS, the spectral regions prospecting changes the window size and moves it all over the spectral regions, highlighting informative regions associated with each window size. Among all possible configurations of window sizes and positions, CSMWPLS selects the most important regions for improvement of state variables prediction by PLS models. In that way, this method optimizes both size and region for variable inference through PLS models.

SCMWPLS it is a natural evolution of CSMWPLS, where not only PLS models of selected regions are obtained, but the best combination of spectral elements

constituents from selected regions to obtain the best group of spectral elements for PLS model prediction of interest state variable. The main idea in CSMWPLS and SCMWPLS is performing combinations of exhaustive search in pre-selected regions by MWPLSR. Once the hole spectral is significantly reduced, these strategies allowed the variable modeling problem resolution in an easy and intuitive way.

Despite their advantages, as allowing the obtention of better PLS models when compared with the whole spectral models, those strategies are tied to combine well correlated regions with the interest variable, in this way; possible effects of interactions between well correlated regions with less significant regions can be masked and neglected. One possible solution can be an association of moving window PLS strategy with a global optimization technique, as Genetic Algorithm (GA), programmed for combinatorial analysis and searching of spectral group elements. In this way, effects of interactions regions can be evaluated, ensuring that less individual significant regions, possibly neglected, have their combined information considered.

This work presents the merger of two chemometric strategies: the changeable size moving windows PLS (and their following modifications) with the pure spectral chemometric modeling (PSCM), and compare both approaches for spectral data pre-selection and variable inference. Once both approaches allows pre-selection of spectral data, and variable infer using multiple input models, all the arrangements of pre-selection and infer methods using those strategies are tested, searching for the best combination for protein quantification on flour samples using NIR measurements.

PSCM is a chemometric strategy presented previously by our group (Ranzan, Strohm *et al.* 2014), where spectral regions are highlighted and spectral elements are selected for chemometric modeling of process variables, using ant colony optimization (ACO) as spectral group analysis tool. The advantage of applying ACO for spectral elements groups remains in the fact that ACO uses a vector for simulate pheromone amount during algorithm routine, and this vector allows qualitative analyses of spectral data.

These features allow the combination of both strategies and create an algorithm for chemometric modeling of spectral data for process variable prediction, where spectral data are pre-selected using CSMWPLS and combined using PSCM.

In order to evaluate the methodologies presented on this study, NIR spectroscopy data from flour samples, previously characterized according their protein content, were used. The selection of this data set is based on the fact that NIR is considered a Standard method for flour characterization (Cocchi, Corbellini *et al.* 2005, Ait Kaddour and Cuq 2009, Li Vigni, Durante *et al.* 2009, Vigni, Baschieri *et al.* 2011), what corroborates it is choose and provides a background for results analysis.

6.2 Methodology

6.2.1 Experimental data set

Data set is compounded by 34 different samples of flour, including wheat flour and rye flour. It was structured in a two dimensional matrix form, where the rows are related to the samples and the columns are related to the measurements, in a way that each

column makes reference to a specific wavelength (spectral component), for all samples. NIR data was collected in triplicate measurements, totalizing 102 spectras Segmentation was made randomly, taking two thirds of samples for the calibration phase, and one third remained for prediction and test phase.

Off-line measurements were performed in laboratory through a complete farinograph analysis, using the Brabender GmbH & Co. KG, Duisburg, Germany, model FD0234H and the digestion apparatus from Digesdahl® Hach - Düsseldorf, Germany, although, only protein content was taken into account. The range of protein content in the sample group varies between six and fourteen grams of protein in one hundred grams of flour sample.

NIR spectral data was collected in a Multi-Purpose NIR Analyzer (Bruker Optik GmbH - Ettlingen, Germany), with a wavelength range from 800 nm to 2800 nm, with a non-continuous increment, leading to spectral information compound by 1050 independent wavelengths.

Real spectral data are full of noise and others nonidealities that mask the data information. Before any chemometric analysis, it is necessary the normalization of the spectral data, allowing the real data can correctly be analyzed. NIR data was normalized using Standard Normal Variate (SNV), which scales the samples instead of the spectral variables (Beebe, Pell *et al.* 1998).

Figure 6.1 presents NIR spectroscopy measurements from the 34 flour samples, here each spectral line correspond to the mean of the triplicate NIR measurements for each sample. Figure 6.1(a) makes reference to the raw NIR data, while Figure 6.1(b) refers to SNV normalized NIR data.

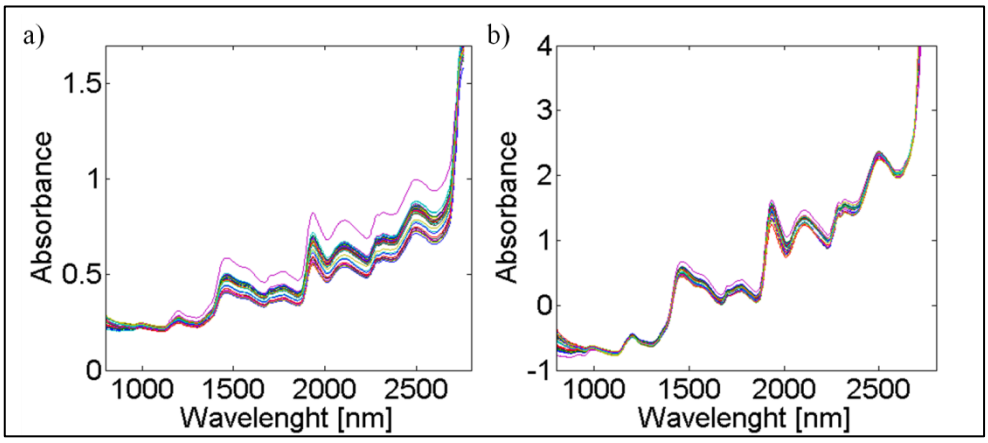

Figure 6.1: NIR measurements average from 34 flour samples. (a) NIR raw data and (b) NIR SNV normalized data.

Qualitative analyses of spectroscopic data are made using principal component analyses (PCA). The graphical visualization of flour samples in function of first five principal components, in the PC's plans, is presented on Figure 6.2, here the mean of NIR triplicate data for each flour sample was evaluated. This figure shows not clear samples group segmentation, although, it is possible conclude that there are some correlation between samples features, especially in function of the flour types (T – wheat flour; C – rye flour).

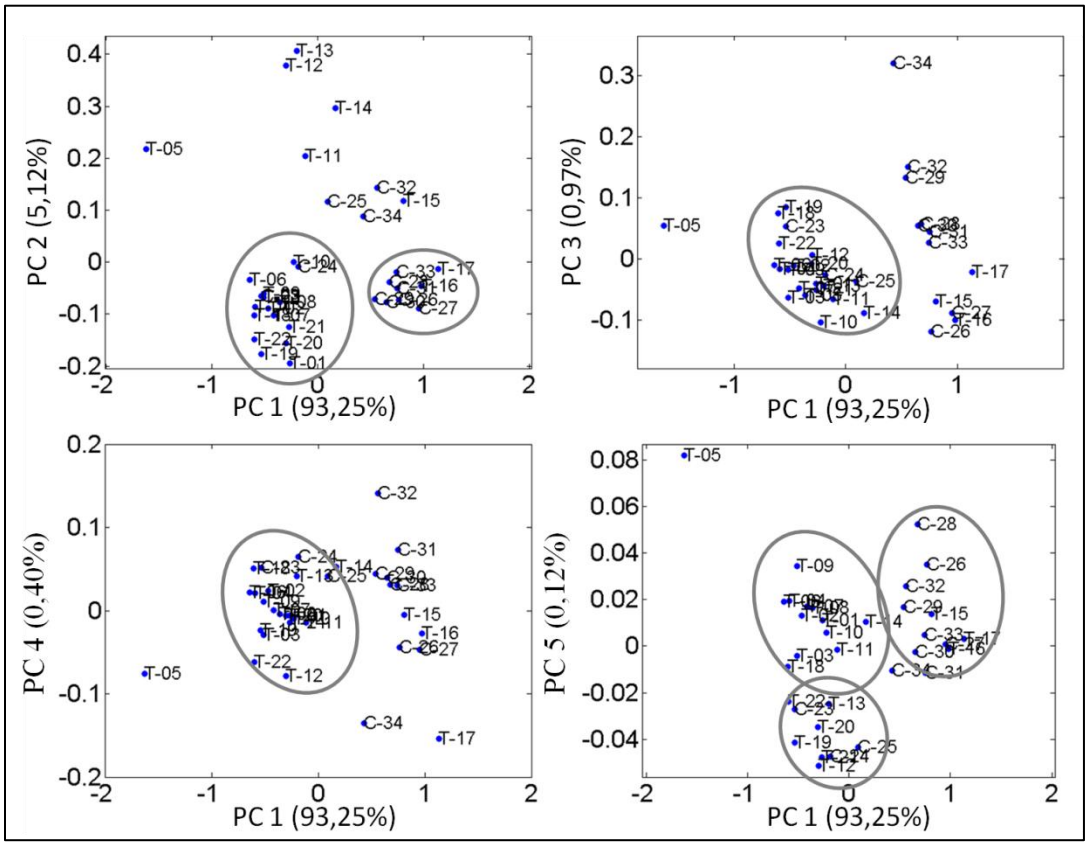

Figure 6.2: 34 flour samples localizations in PC's plans, using PCA results for qualitative data set evaluation (T – Wheat Flour; C – Rye Flour).

6.2.2 Chemometric Methods

All routines was implemented in MATLAB software (Ver. 5.3.0.10183 R11, The Mathworks, Inc., Natick, USA).

The proposed models were compared according the statistical parameter Root Mean Squares Error (RMSE), for accuracy ranking. This parameter is named differentially when calculated in the calibration or prediction phase: RMSEC for calibration and RMSEP for prediction. In equation 6.1 the sub index p refers to the vector of the predicted variable values, resulting from model evaluation, and m to the measured variable values, N is the number of measurements and y is the vector of interest variable.

$$\boldsymbol{RMSE} = \sqrt{\sum_{i=1}^{N}\left(y_{p,i} - y_{m,i}\right)^2 \Big/ N} \qquad (6.1)$$

Moving Window Partial Least Squares Regression - MWPLSR

Initially presented by Jiang *et al.* (2002), MWPLSR searches for the significant spectral regions by scanning spectral data with a partial least squares regression procedure over continuous sweeping analysis of spectral intervals. In MWPLSR, a sequence of PLS models

is built for a moving window through the whole spectra. Qualitative mapping of spectral data is obtained in terms of the decrease of prediction error, showing an improvement of PLS models predictions (Du, Liang *et al.* 2004).

The kernel of MWPLSR rises on a moving window. This window is made by a certain number of spectral elements, defined by the user, and called window size (h). It starts at ith spectral element and ends at (i + h +1)th element, comprehended by all the elements inside this interval. For each window size, there are n – h +1 windows over the whole spectra, where n is the number of spectral elements available at the full spectral data. Each window is considered a subset of original data and PLSR models are generated for each window. The number of load vectors (LV) used for modeling is determined by the user, although, its maximum number should be equal or smaller than the lowest dimension of the subset data matrix (number of calibration samples x h).

The prediction was evaluated by the sums of squared residues (SSR) presented by each window position. The value of SSR in function of window position is plotted for the first element constituent of the subset, and its behavior in function of the spectral region is evaluated proportionally for the range of SSR. Regions with significant dependence of the interest variable show smallest values for SSR. In that way, valleys in the curve of SSR in function of window position indicate regions of significant information content (Du, Liang *et al.* 2004)..

Changeable Size Moving Window Partial Least Squares - CSMWPLS

CSMWPLS is a method for optimization of informative spectral regions, preselected by MWPLSR, focusing in search of spectral sub-regions using different window sizes, in the opposite of MWPLSR, here only one window size is scanned. This method was presented by Du *et al.* (Du, Liang *et al.* 2004) as a way to refine the search of significant spectral regions and improvement of PLS models predictions.

The basic idea of CSMWPLS is to move many windows inside a defined spectral region, varying the window size from one until the size of the spectral region defined. During this process, all the sub-windows with the same window size (h) are obtained. For every window, PLS models are obtained, using the number of LVs previously determined as the most indicated, and RMSEC (root mean square error or calibration) values are used for spectral region refining. Figure 6.3 presents a graphical representation of MWPLSR and CSMWPLS methodologies, showing the sliding window inside the spectral dimension. For each position of the window, a new PLSR model is generated using calibration data set and tested with prediction data set. The result of RMSEP for each window is associated with the first element comprehended by the window.

Once CSMWPLS is an extension of MWPLSR, results are compared for qualitative and quantitative analysis. The main idea of spectral region selection using CSMWPLS is search for a combination of window size and position inside spectral range that presents smaller values of RMSEP.

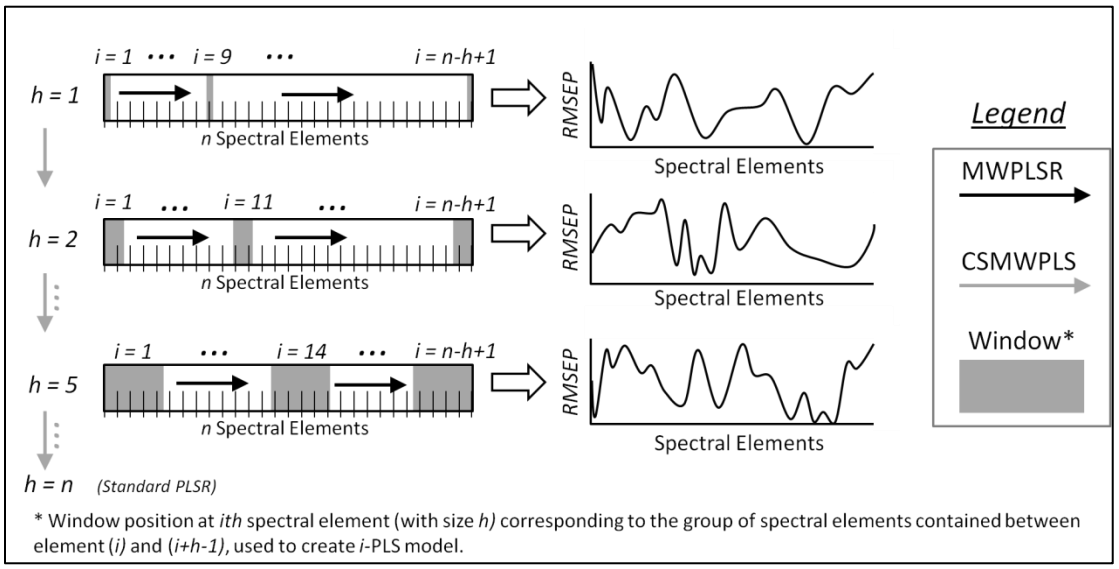

Figure 6.3: Schematic representation of MWPLSR and CSMWPLS.

Pure Spectra Chemometric Modeling – PSCM/ACO

This chemometric analysis has presented by our group in a previously work (Ranzan, Strohm *et al.* 2014), where is described in details the two main pillars of this method: the selection of pure spectral elements and model adjustment for state variable prediction.

The steps for chemometric modeling using the PSCM methodology can be divided into three phases: Selection of spectral elements group; models calibration, and models validation. The selection of the spectral group aims to choose spectrum components that when combined using MISO (Multiple Input Single Output) models allow direct correlation with state variables (Skoog, Holler *et al.* 2007). Those models have the characteristic of being linear in relation to their parameters, allowing the estimation of its parameters using ordinary least squares with the analytical solution (Joe Qin 1998). Once PSCM is a supervised learning method, the quality, representativeness and the number of measured points have a strong impact in the model quality. Thus, we seek the best, largest and most representative data set to build the models.

The selection of the independent variables to be included in the spectral component set is the most important step to build a PSCM. For that, in this work, we apply the Ant Colony Optimization (ACO) method. The benefits obtained from this optimization approach are the stability of the model in terms of collinearity in multivariate spectra and the interpretability of relationship between spectral data and sample compositions as initially shown by Allegrini and Olivieri (2011).

ACO is based on real ants behavior, more specifically, by the indirect communication between them within the colony using chemical pheromone secretion (Dorigo and Blum 2005, Dorigo, Birattari *et al.* 2006, Mullen, Monekosso *et al.* 2009). The main idea behind this algorithm is that, in real ants, the convergence of ant trails toward the shortest route between the food source and the nest is a result of the tendency of ants to follow a trail that contains a higher concentration of the pheromone deposit (Deneubourg, Aron *et al.* 1986). Details about Ant Colony Optimization algorithms and its implementation can be obtained in Mullen *et al.* (2009) and Ranzan *et al.* (Ranzan, Strohm *et al.* 2014).

A great advantage associated to PSCM is the capability of the method to characterize the spectral data in a qualitative way, based on the pheromone vector, created during ACO running. This vector works as independent numerical information which ants use probabilistically to construct solutions to the problem. Its updates occur after each interaction of ACO and contain information on each spectrum component's relevance to state variable prediction. A previous work (Ranzan, Strohm *et al.* 2014) showed that the application of PSCM and ACO for spectral regions selection can improve PLS models variable prediction around 60% for the coefficient of determination, when compared with standard PLS models (without spectral selection).

6.2.3 Chemometrical approaches arrangement

CSMWPLS and PSCM/ACO are both chemometrical approaches with application on pre-selection of optimized spectral data for variable infer, and for obtaining of chemometrical models for transduction of spectral data into variable information. Each one of those methods has some advantages and disadvantages, more or less indicated for each process.

This work evaluates all the possible combinations of these two chemometrical methodologies, applied for conversion of NIR spectral data into protein content in flour samples. CSMWPLS and PSCM/ACO were alternated in pre-selection of spectral regions and variable modeling. Results are compared with standard PLS, CSMWPLS and PSCM/ACO strategies applied to full spectral data. In this way, the best combination between both methods is obtained for flour samples characterization. We also obtained the indication of the best features associated with each method.

6.3 Results and Discussions

6.3.1 Standard CSMWPLS

Wheat flour SNV normalized data was previously submitted to a standard PLSR modeling, providing the basis for results comparison. The PLSR analysis was made using the full spectral data, without any pre-selection of spectral regions, and varying the number of load vectors (LV) used on models from one to twenty. Figure 6.4 shows the graphics of explained variance presented by the first twenty LVs and the result of RMSEP for protein prediction presented in function of LVs numbers used on models development.

The results of PLSR standard analysis indicate that the best number of LVs to use in a PLS models, for this specific data set, are around 11 and 13, since on these models the minimum of RMSEP was achieved, with values of explained variance reaching more than 95%.

After initial PLSR modeling, CSMWPLS was applied for spectral data classification. In the first run of CSMWPLS, the window size varied from one to four hundred with simultaneous variation in the LVs number used on the models from two to twenty. In that way, were searched for the optimized combination of single window size (single spectral

region) and PLS number of components (LVs). The best results of RMSEP in function of LVs are presented in Table 6.1, indicating the optimized spectral region used in each prediction model. The spectral regions of Table 6.1 are plotted in Figure 6.5, allowing the visualization of spectral regions selection by CSMWPLS, as a function of PLS models size.

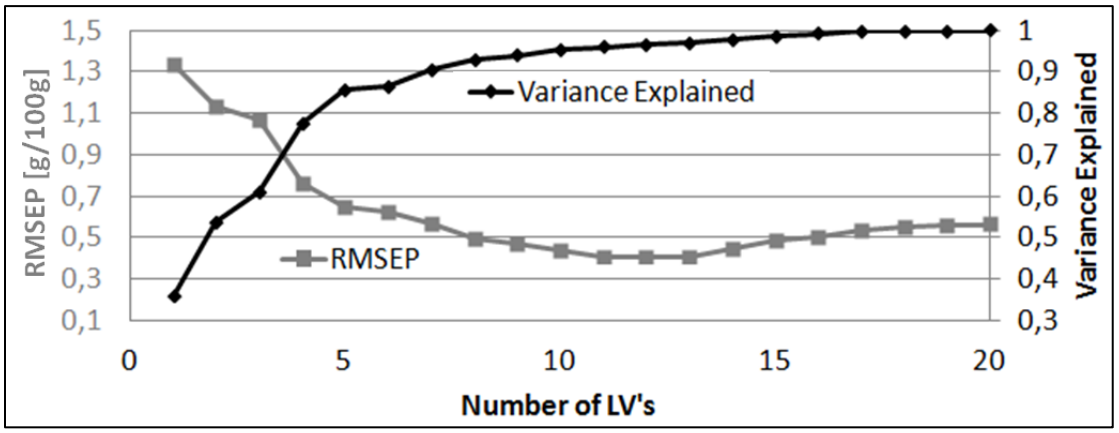

Figure 6.4: Results of RMSEP and Explained Variance for PLSR modeling applied to full spectral data of wheat flour and respective protein content.

Tabel 6.1: CSMWPLS results of RMSEP for the best obtained PLS models

Number of LV's	RMSEP [g/100g]	Wavelength Interval [nm]	
2	0.656	2035.10	2080.83
3	0.595	2135.68	2215.99
4	0.513	1168.94	1289.91
5	0.492	1158.50	1301.56
6	0.449	1143.17	1418.33
7	0.434	1109.89	1399.95
8	0.388	1108.95	1402.98
9	0.392	1109.89	1402.98
10	0.398	1087.55	1510.91
11	0.421	1662.00	2383.01
12	0.423	1664.13	2387.40
13	0.430	1640.96	2369.94
14	0.437	1703.49	2436.76
15	0.437	1705.73	2432.19
16	0.448	1640.96	2432.19
17	0.483	1628.59	2357.01
18	0.487	997.20	1432.44
19	0.487	997.20	1432.44
20	0.487	997.20	1432.44

,

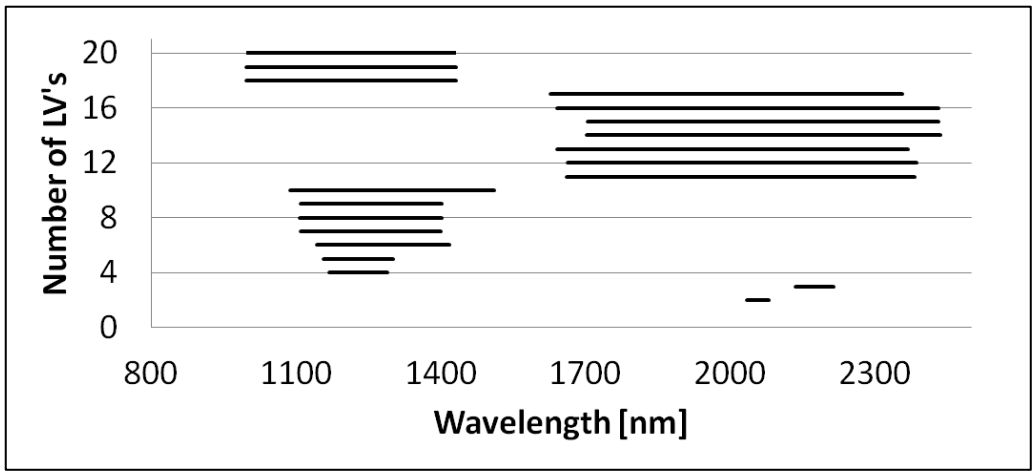

Figure 6.5: Optimized regions of NIR spectra for prediction of protein in function of LV's number on PLS models, applying CSMWPLS.

The results reported in Table 6.1 and Figure 6.5 show an improvement in the prediction using PLS models. The best result of CSMWPLS, using a single spectral region was obtained using a model compounded by eight LVs and using the NIR spectral region comprehended between 1109nm and 1403nm. PLS models with nine LVs, with spectral region similar to the previously one, and ten LVs, with a spectral region around 40% larger than the previously indicated regions, have similar values of RMSEP. The best results obtained with CSMWPLS presented better protein prediction when compared with standard PLS models. The smallest RMSEP achieved using all spectral range was equal to 0.41 (see Figure 6.4, for 10, 11 and 12 LVs models), while CSMWPLS obtained a PLS model able to achieve RMSEP values of 0.388 using less number of LVs and only 21% of the total amount of spectral elements available.

Results presented by Du *et al.* (2004) and Jiang *et al.* (Jiang, Berry *et al.* 2002) indicated that the classification of spectra regions was dependent of PLS models size, which variation implicates in resolution of regions degree of differentiation, but not in regions changes. In other words, for a continuous window size, the variation of PLS models sizes implies in increase of RMSEP valleys and consequently better segmentation of significant spectral regions.

Once the best PLS model obtained from previously CSMWPLS running was found with PLS model size of 8 LVs, evaluation of subsequent spectral filtering was developed with this model size, keeping it constant.

The selection of the spectral region using usual CSMWPLS could not be made comparing the results of RMSEP from different window size. As showed on Figure 6.6, the increase on window size, keeping the PLS models size constant, implies in the displacement, in the direction of small wavelengths of the peaks and valleys of RMSEP. The changes of highlighted spectral region become a problem when different windows size results are compared.

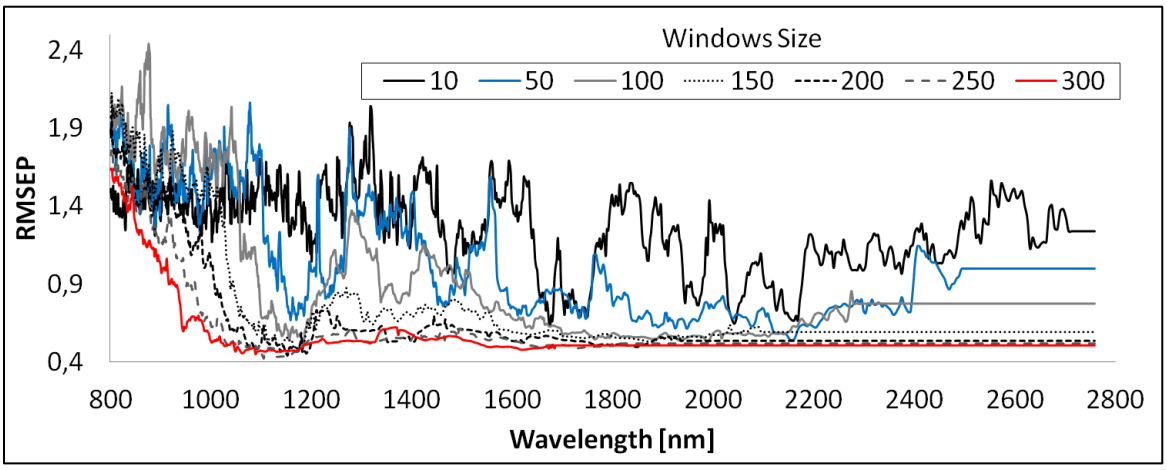

Figure 6.6: RMSEP values for protein prediction obtained using CSMWPLS for NIR spectral region selection, for seven distinct windows size: 10, 50, 100, 150, 200, 250 and 300 spectral elements window**.**

6.3.2 modCSMWPLS

To solve this segmentation region problem, we are proposing a modification on the original CSMWPLS (modCSMWPLS). The method CSMWPLS proposes for spectral region segmentation does the individual analyses of RMSEP for each window size. It does not evaluate the effect of each spectral element as unique but only in the group. The modification proposed in this work intends to evaluate each spectral component contribution in the efficiency of process variable prediction.

Inspired by the pheromone trail evolution during ACO routine (Ranzan, Strohm *et al.* 2014), CSMWPLS algorithm was modified by the insertion of a vector for RMSEP storage, in function of spectral elements applied on PLS models and state variable prediction. This vector intends to evaluate the prediction capability of models by the insertion of spectrum elements on it, in a generalized way, not been dependent of the windows size for region highlighting.

CSMWPLS routine can be partitioned in subsequent MWPLSR runs. For each window movement, during a single MWPLSR search, a PLS model is obtained and tested for variable prediction, using the set of spectral elements contained in the window. The vector added on CSMWPLS routine, stores, on respective positions for spectral elements in the analysis, the value of RMSEP for the tested PLS model, adding the result to the previously storage values. This approach is repeated, subsequently, for all window sizes selected by the user, and between the window transitions, each element on storage vector is divided by the number of times that its correspondent spectral element was used on a model group. The vector of summed RMSEP results is initialized with zeros in the beginning of modCSMWPLS routine.

Figure 6.7 presents the results of RMSEP storage during modCSMWPLS, varying window size from one up to fifteen. The position of valleys and peaks are constant, improving the process of spectral region selection. Once the storage vector decreases its variation with the addition of new windows search steps, it is crucial the right combination between quantities of window search take into evaluation, and the desired

degree of region segmentation. As higher the amount of window searched, smaller is the differentiation between reliable and not reliable regions for variable prediction.

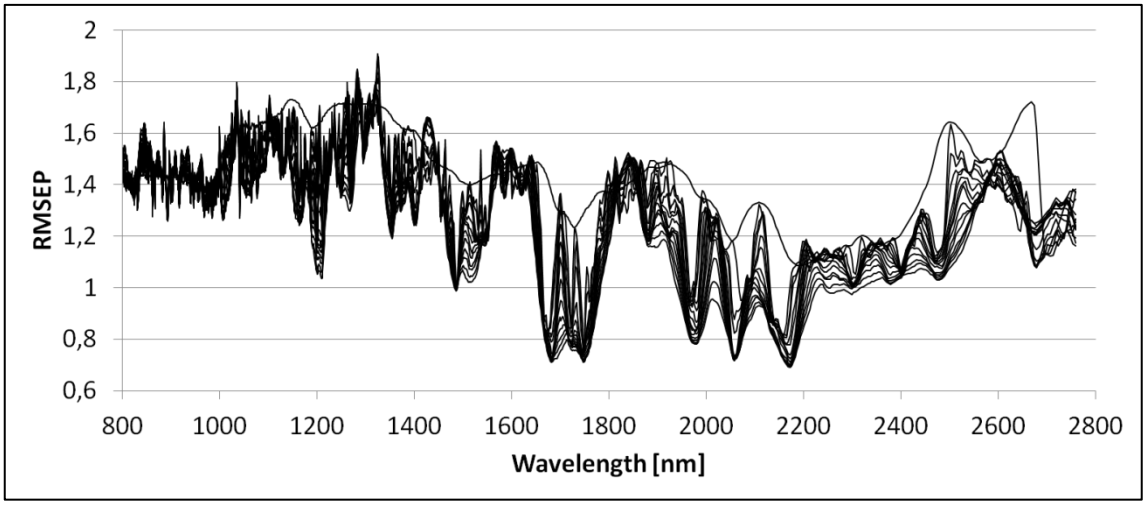

Figure 6.7: RMSEP storage values for modCSMWPLS, with window size from one up to fifteen, applied for NIR data of wheat flour samples and using PLS models with maximum of 8 LV's**.**

6.3.3 Chemometrical Approach – Filtering and Modeling

Results shown in Figure 6.7 indicates, approximately, four NIR spectra regions that present higher correlation with protein content in wheat flour samples (visual inspection). Those regions are comprehended between 1140nm – 1270nm, 1450nm – 1550nm, 1650nm – 1780nm and 1950nm – 2200nm. Each highlighted region can be directly associated with knowledge about NIR overtones of flour samples. The regions indicated by modCSMWPLS are associated with C-H stretch, O-H stretch, C-H stretch and amide overtones, respectively.

For comparison of spectral data selection, results of modCSMWPLS for protein characterization in wheat flour samples are confronted with PSCM/ACO results for the same purpose. PSCM/ACO strategy was applied analyzing pheromone trails from one until nine pure spectral elements, and each pheromone trail was normalized for mean comparison. Figure 6.8 presents the highlighting of NIR spectral region presented by PSCM/ACO for the experimental data set.

Results presented on Figure 6.8 are related with the pheromone amount deposited by ants on spectral elements, it is a qualitative criterion for selection of elements into PSCM models, and it relative higher amounts are connected with better elements for process variable prediction.

Some spectral regions highlighted by PSCM/ACO are similar to the ones presented by modified CSMWPLS, as the regions comprehended in 1140nm – 1270nm, 1650nm – 1780nm and 1950nm – 2200nm. Despite those regions, PSCM/ACO also indicates the region comprehended between 2550nm – 2690nm as significative for protein prediction. Those regions were not selected by CSMWPLS. There was a valley around 2690nm, although, it did not show significant RMSEP to be included into the selected regions.

Both approaches for spectral data selection emphasized NIR spectral regions associated with knowledged overtones. According to the works of Sun (Sun 2008) and Champe and Harvey (Champe and Harvey 2005) the regions indicated by pre-selection methods as significant for protein inference are high correlated with C – H stretch bands (overtone of 1200nm and 1800nm) and amide band (2100nm). In our previously work (Ranzan, Strohm *et al.* 2014) we correlated the spectral regions highlighting for PSCM/ACO with interesting variable features expected into NIR spectral overtones associated with protein samples. The results showed equivalent results to the ones obtained by modified CSMWPLS (Figure 6.7) and PSCM/ACO (Figure 6.8) spectral regions selection.

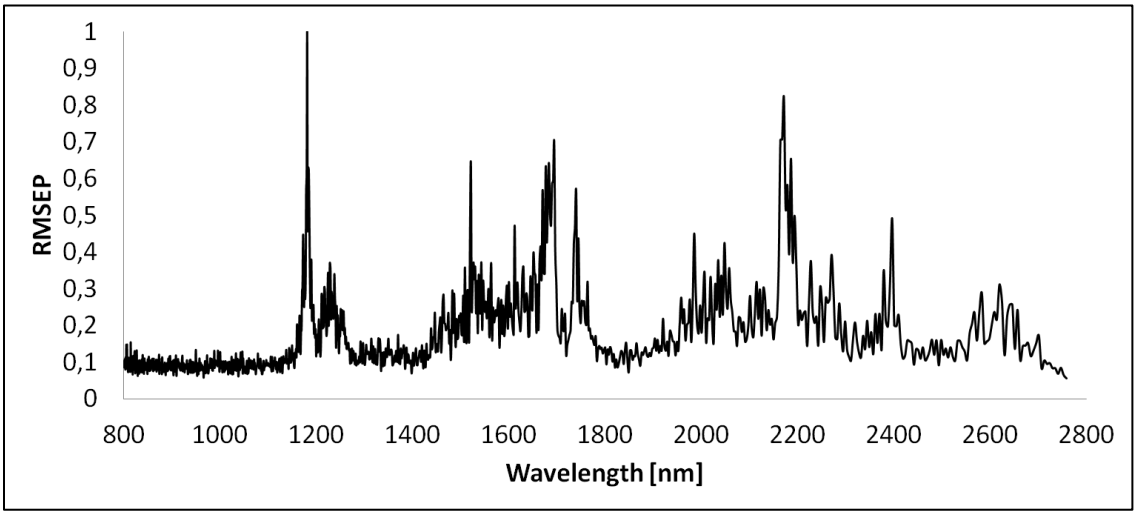

Figure 6.8: Relative pheromone amount deposited in spectral elements, obtained using PSCM methodology varying models from one up to nine spectral elements prediction groups.

The qualitative differences between Figures 6.7 and 6.8 are mainly results from the approach used in each spectral search. While CSMWPLS (Figure 6.7) bases its spectral analyses into PLS modeling using sequential spectral elements, PSCM/ACO searches for non-sequential spectral groups. The first method is limited by the fact that if single wavelengths are highly sensitive for the interested compound, but are located in a spectral region where the other elements are not, CSMWPLS will have a small valley at that region, and could mask the important information contained on single elements. This disadvantage is not present on PSCM/ACO, once this method searches for spectral groups in an individualized approach, and the importance of each spectral element, for variable prediction, is evaluated separately, even through models propositions and tests are made in groups.

The results of spectral regions selection from modCSMWPLS and PSCM/ACO are compared using tree chemometric methodologies: PLS; standard CSMWPLS, where best results of PLS models with variable windows size are tested; and PSCM/ACO methodology, where filtered spectral data is combined individually, using multilinear models. In this phase, the selected regions are used for protein content inferring.

The results of protein prediction using both pre-selected data process are shown in Figure 6.9. The two new spectral data set are compounded by same amount of spectral

elements, corresponding to 20% (230 elements) of the total amount (1150 elements). In both filtering methodologies, the selection of spectral elements is divided into 2 phases: selection of spectral significant regions, as described previously for each approach, and sorting of spectral elements inside each region, based on the qualitative information of each approach. For CSMWPLS, the sorting is based on the lowest summed RMSEP values, inside each region separately, while in PSCM/ACO the selection of spectral elements is made considering the highest values of pheromone in the total spectrum data range, as once.

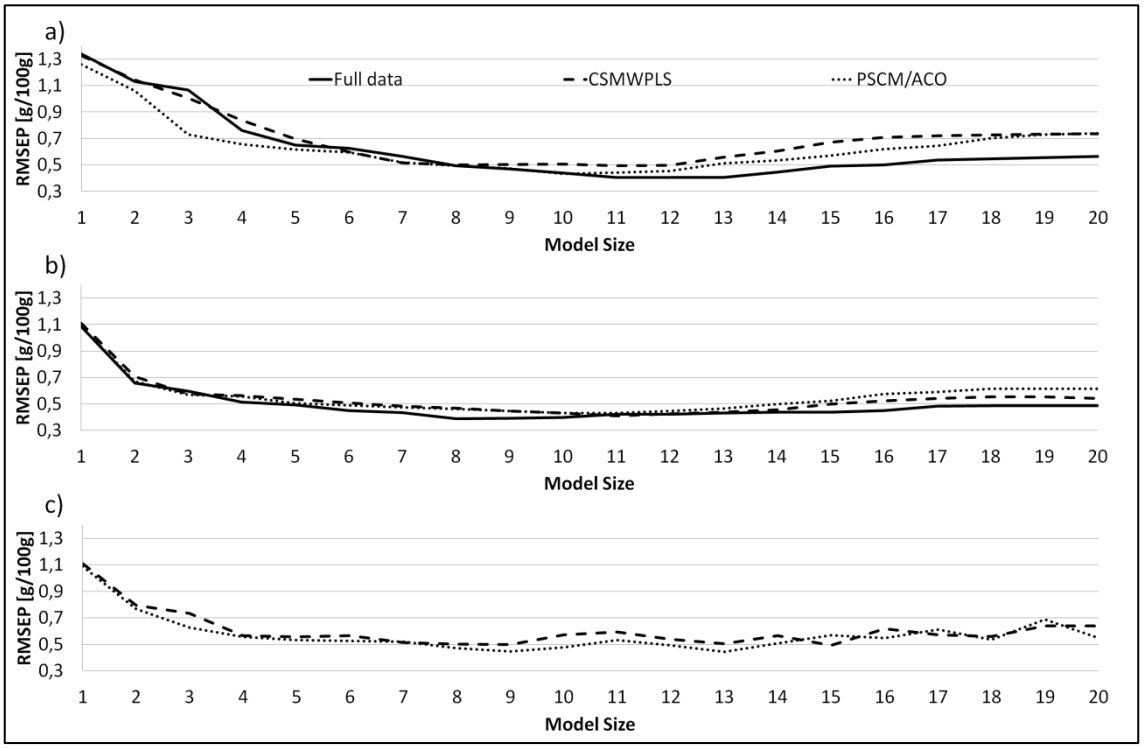

Figure 6.9: RMSEP results for chemometric modeling of protein content prediction using NIR data from wheat flour samples, using full spectrum data, filtrated spectrum data using modified CSMWPLS and filtrated spectrum data using PSCM/ACO. Results are presented in function of independent variables used on chemometric models. Modeling process divided into (a) standard PLS regression, (b) CSMWPLS and (c) PSCM/ACO.

Figure 6.9(a) presents RMSEP modeling results of standard chemometric methodology PLS applied on full spectral data, in comparison to PLS models results using filtered spectral data. Both filtrated data provides similar results of RMSEP in function of PLS models size, indicating that both filtering methodologies can effectively condense useful information in less number of elements. Results of PLS models using PSCM/ACO filter presented better values of RMSEP for PLS models with less LV's, although, this characteristic is not constant when increasing LVs quantity. The best result for protein prediction obtained by PSCM/ACO filtered data using PLS models is obtained using 10 LV's, while, for filtered data using modified CSMWPLS 8 LVs were enough. Even presenting similar results, both models achieved higher RMSEP indices (46% and 23.3%, respectively) in comparison with the best result using full spectral data.

Results presented on Figure 6.9(b) and Figure 6.9(c) are related with protein prediction achieved using the same methodologies applied for spectral filtering. Both

results are used for evaluation of prediction capability of filtered spectral data, using different approach than standard PLS models. Those results confirm the condensing capability presented by discussed methods, and shows that spectral data selected using PSCM/ACO produced better results for lower models sizes than using standard CSMWPLS modeling, indicating that the filtering procedure applied by PSCM/ACO is more efficient in spectral information summarization.

An interesting point is that both spectral filters organize spectral data filtered in wavelength order. The position of selected regions could influence CSMWPLS results, as discussed by Du, *et al.* (Du, Liang *et al.* 2004), although the evaluation of elements optimum position can be as hard as spectral filter implementation, since the redistribution of spectral data for CSMWPLS modeling improvement becomes another optimization and spectral selection problem.

For better comparison between predictions results showed in Figure 6.9, Figure 6.10 presents the percentage difference of RMSEP results between RMSEP values of PLS models using full spectral data and the models obtained using PLS, CSMWPLS and PSCM chemometric methodologies and the different spectral data selected. On Figure 6.10, PLS using full spectral data curve correspond to the line of zeros. In that way, the respective curve of each modeling procedure and its respective spectral data set is directly associated with the zero line, concluding that higher the values assumed by the curves, better is the prediction presented by the model in comparison to standard PLS models. These results are made reducing from the RMSEP values obtained by standard PLS the values presented by each chemometric methodology and dividing by the value of standard PLS.

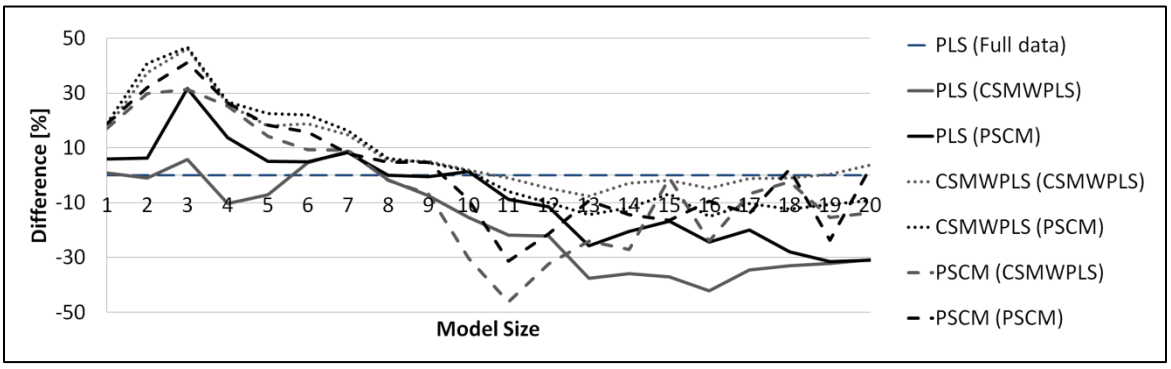

Figure 6.10: Percentage difference between RMSEP results for protein prediction of PLS chemometrical models using full NIR spectral data and: PLS models using NIR filtered data with modified CSMWPLS (PLS(CSMWPLS)) and PSCM/ACO (PLS(PSCM)), CSMWPLS models using NIR filtered data with modified CSMWPLS (CSMWPLS(CSMWPLS)) and PSCM/ACO (CSMWPLS(PSCM)) and PSCM/ACO models using NIR filtered data with modified CSMWPLS (PSCM(CSMWPLS)) and PSCM/ACO (PSCM(PSCM))

Results presented on Figure 6.10 corroborate the conclusions from Figure 6.9, where models obtained using filtered data results in smaller RMSEP values for models with less than 8 input variables. Filtered spectral data obtained using CSMWPLS have the better results in comparison with standard PLS and full spectral data, especially for models with 3 input variables, achieving improvement in RMSEP values in order of 46%.

Chemometric models with more than eight input variables have a decrease on prediction capability using filtered data, although, as showed in Figure 6.9, models with more than 8 input variables, presented higher values of RMSEP, indicating that the maximum models size for filtered spectral data for wheat flour characterization is equal to eight.

The best result for increase of PLS models was obtained using PSCM/ACO strategy for NIR data selection. In function of models size, the use of pre-selected data obtained by this approach leads to a increase of 30% into model accuracy, using 3 input variables.

6.4 Conclusions

The huge amount of spectral information provided by optical measurement techniques and high sensitivity and correlation between spectrum components hinders our task of transforming information into process knowledge. Nonetheless, the significant amount of noise and medium complex characteristics makes the application of multivariable modeling tools a challenging problem.

Spectral pre-selection methodologies are a significant subject for spectral data pre-treatment, blocking spectrum components with low correlation to the inferred variables and improving the model robustness and quality. Also allows the non full range spectral measurements, reducing significantly the number of collected points and information for storage and/or to process.

This paper compares different approaches for pre-selection of NIR spectral data. The protein content prediction in wheat and rye flour samples using these techniques are compared with the standard PLS chemometric method using the full range spectral data.

Well known chemometric modeling approach MWPLSR and a methodology extension, CSMWPLS, can be used as tools for spectral data filter, basing its selection into multiple PLS models, applied to segmented spectral range into various spectral windows. Despite the innumerous advantages of this method, it has the disadvantage of changing qualitative characterization of spectral data with the variance of windows spectral segmentation. To solve this problem, in this work, it was proposed a change into the evaluation of spectral data adding a vector of summed RMSEP values for subsequent window evaluation. This approach was based on ACO pheromone trail actualization.

The modified method provided good results for condensing NIR spectral data from wheat flour samples, tested for reduced total spectral data into 21% of its all spectral components, and been efficient to maintain the main elements, corresponding to the ones responsible for protein content prediction on samples.

Comparing the spectral filter based on the modified CSMWPLS with the PSCM/ACO approach, we can conclude that both methods have similar results in the selection of spectral regions for the experimental data set, indicating the viability in application of those techniques in spectral data treatment. Despite the good results presented by both approaches in comparison to full spectral data, PSCM/ACO achieved smaller results of RMSEP, indicating its better data condensing capability.

For the most chemometric models tested (i.e., PLS, CSMWPLS and PSCM) pre-selected data presented smallest RMSEP values than the same size and model type using full spectral data. Usually for chemometric model prediction, the model order has an optimum size, since for larger models prediction becomes less accurate. Filtered spectral data reduces the size and improve the data quality, what can improve the final model quality.

For all tested chemometric approaches, PSCM models shown better results in function of models sizes, leading to the conclusion that this method is more appropriate to improve the model quality than PLS or even CSMWPLS. The proposed modified CSMWPLS has also shown a very good spectral condensing capability, allowing more accurate models adjustment with more than 45% of accuracy improvement using 3 input variables, in comparison with standard PLS modeling and full spectral data

Capítulo 7 – Conclusão e Trabalhos Futuros

O tratamento de dados de espectroscopia se mostrou complexo, mas de grande aplicabilidade no universo industrial, com grande ênfase nas áreas correlatas à engenharia química. O déficit de metodologias apuradas para caracterização de processos é um problema que vem sendo pesquisado, principalmente por profissionais da área de controle e otimização de processos, já que a capacidade de acompanhar as variáveis de interesse impacta de forma direta na aplicabilidade das estratégias de controle propostas.

Neste âmbito, a pesquisa e desenvolvimento de sensores de processo, baseados em medidas óticas, surge como a evolução natural para a proposta de sensores personalizados para processos. Sensores óticos possuem inúmeras vantagens associadas à quantificação e transdução de informações a partir de ondas eletromagnéticas, características essas, que se adéquam perfeitamente na grande maioria das situações nas quais sondas e sensores estão sujeitas. Além das vantagens práticas apresentadas, sensores baseados em medidas óticas ainda possibilitam a coleta de dados com alta precisão e acuracidade, caso projetados de forma adequada.

A estratégia de transdução das informações contidas em conjuntos de dados espectrais em conhecimento útil acerca de processos, apresentada neste trabalho, PSCM, mostrou-se de grande valia na área da quimiometria, resultando em modelos com capacidade preditiva equiparada àquelas obtidas com a utilização de metodologias quimiométricas consolidadas desta área de atuação.

Apesar da grande dificuldade de converter informação em conhecimento, a aplicação de ferramentas de modelagem multivariada assim como PCR e PLS, obteve bons resultados e mostrou grande aplicabilidade, baseado nos estudos desenvolvidos com ambos os conjuntos de dados experimentais abordados neste trabalho.

Quando são comparadas as capacidades preditivas de inferidores de grandezas de interesse ajustados usando PCR e PLS, para predição de composição de proteína em amostras de farinha de trigo e centeio e utilizando dados de espectroscopia NIR, ambas apresentam eficiência equivalente. Entretanto, como esperado, modelos PLS atingiram melhores resultados, com um número reduzido de variáveis independentes (vetores de

carga), levando a concluir que inferidores de estado propostos para este sistema, utilizando a metodologia PLS, são os mais indicados.

Apesar dos resultados satisfatórios apresentados pelos modelos PLS, melhores resultados foram obtidos com a utilização da metodologia PSCM. Inferidores ajustados com este método obtiveram maior precisão na predição do conteúdo proteico das amostras do conjunto experimental. Este resultado confirma que a seleção e combinação de elementos espectrais puros apresentam vantagens em comparação aos métodos quimiométricos que condensam informações a partir de combinações lineares dos elementos, como é o caso dos métodos PCR e PLS.

Além das vantagens apresentadas no quesito modelagem, a metodologia PSCM associada ao algoritmo ACO apresenta a grande funcionalidade de fornecer dados qualitativos para o mapeamento do espaço espectral em função das variáveis de interesse do processo.

A vantagem associada à caracterização qualitativa dos dados espectrais transcende a utilização desta como fonte auxiliar de informação, mas permite a realização de tratamento efetivo de dados em função da variável de interesse. A partir de dados pré-selecionados, inferidores de estado otimizados podem ser ajustados, tornando as predições mais precisas e menos influenciáveis por flutuações no meio reacional. Estas características associadas ao método PSCM mostram o quão viável é esta metodologia para a caracterização de processos de uma forma geral, seja na proposta de inferidores seja como fonte de conhecimento sobre processos.

Quantitativamente, dados de espectroscopia NIR filtrados através da metodologia PSCM propiciaram incrementos na capacidade preditiva de modelos PCR e PLS, aumentando a qualidade de predição (de acordo com o índice R^2) em cerca de 60%, para os melhores casos.

Uma vez que a filtragem de dados espectrais é um assunto de significativa importância na área de quimiometria, foi feita a análise da viabilidade de combinar as funcionalidades das metodologias PSCM com a conhecida técnica CSMWPLS. Ambas as técnicas possibilitam o ajuste de modelos quimiométricos além de servirem como base para filtragem de dados, cada uma com características e vantagens específicas.

Neste contexto, diversos testes de associação entre as duas metodologias foram realizados, combinando ambas de forma alternada para filtragem e modelagem do conjunto de dados de farinhas. A taxa de filtragem utilizada foi de 79%, de forma que apenas 21% do total de informação espectral contida nas medidas NIR foi mantida.

A versão de CSMWPLS utilizada neste estudo foi modificada pela introdução de um vetor para quantificação da qualidade de predição associada a cada elemento espectral (baseado no vetor de trilha de feromônios do algorítmo ACO), melhorando significativamente a resolução qualitativa dos dados espectrais, frente a variação do tamanho de janela avaliado. A esta versão modificada, deu-se o nome de modCSMWPLS.

Na predição do conteúdo de proteínas presente nas amostras de farinha, para todas as combinações de metodologias, a utilização de dados filtrados resultou em melhores predições, frente modelos PLS padrão. A melhor capacidade preditiva foi obtida com dados filtrados utilizando a metodologia PSCM e modelados com a metodologia

CSMWPLS. Apesar deste resultado positivo, modelos PLS usando a matriz completa de dados NIR apresentaram resultado geral melhor, atingindo menor valor de RMSEP específico. Apesar disso, para modelos com menos de 8 variáveis de entrada, dados filtrados, independentemente do método de filtragem ou ajuste de modelo, apresentaram menores valores de RMSEP.

Com relação à caracterização de bioprocessos, a predição de variáveis de estado do processo fermentativo de glicose com a levedura *Saccharomyces cerevisiae* mostrou resultados promissores com a utilização da técnica de espectroscopia de fluorescência 2D.

Inicialmente, para a realização do procedimento de caracterização com as técnicas em questão, é necessário o conhecimento dos referidos estados nos respectivos intervalos de amostragem, e para isso, a simulação de modelo dinâmico, ajustado para o sistema é o método mais indicado. Para isto, foi proposto um novo modelo de taxa de crescimento, capaz de simular de forma adequada, o comportamento diáuxico apresentado pelo micro-organismo, em fermentações batelada. A nova estrutura de taxa de crescimento simula a variação da fonte de substrato consumida em função da concentração de glicose no meio fermentativo.

De posse dos estados caracterizados, foi feita a avaliação da capacidade preditiva de inferidores de estado para este sistema, ajustados com os métodos PCR, PLS e PSCM. Os resultados de predição dos três estados de interesse (concentrações de biomassa, etanol e glicose) obtiveram altos valores de R^2 e baixos valores de RMSEP, mostrando a aplicabilidade desta técnica espectroscópica neste processo.

A avaliação do conjunto de dados deste sistema, utilizando PSCM, permitiu o mapeamento das regiões espectrais com maior correlação com cada um dos estados. Estes resultados possibilitam a obtenção de um melhor entendimento a respeito não só da proposta de futuros sensores, mas também do processo como um todo.

De forma geral, pode ser concluído com este trabalho que a metodologia baseada em componentes espectrais puros, PSCM, apresenta viabilidade de aplicação no estudo e caracterização de dados de processos industriais, principalmente com sua utilização conjunta com a ferramenta de otimização ACO. A aplicação desta estratégia de segmentação dos dados espectrais em seus elementos constituintes permite a construção de sensores ajustados para analitos específicos e mensurados através de informações distintas das regiões espectrais. O fato de combinarmos mais de uma região espectral em modelos com baixo número de variáveis de entrada, menos influenciáveis por ruídos e variações no meio reacional.

A estratégia de estudo apresentada neste trabalho, além de fornecer uma ferramenta prática e rápida para avaliação de dados de processos, também abre a possibilidade de obtenção e caracterização de regiões espectrais com relação às variáveis de estado, e não somente com relação a analitos, como encontrado atualmente na literatura.

7.1 Sugestões para trabalhos futuros

A metodologia apresentada e desenvolvida neste trabalho para o estudo e caracterização de dados espectrais para posterior aplicação na predição de estados de interesse em processo, se mostrou satisfatória com relação a transdução de informações espectrais em conhecimento de processos. Apesar disso, diversas melhorias desta estratégia podem ser vislumbradas, além de novos campos de aplicação para a mesma.

Dentre as modificações que podem ser testadas na metodologia PSCM/ACO, está a seleção de variáveis espectrais utilizando diferentes métricas estatísticas como fatores de otimização. Na implementação amplamente testada neste trabalho, o otimizador está baseado na minimização do erro quadrático médio para a etapa de calibração. Uma possível variante, e que pode vir a gerar bons resultados, é utilizar a maximização do índice de correlação entre os dados preditos e medidos para o conjunto de calibração. A utilização de um indicador estatístico diferente pode reduzir a convergência do método de otimização em mínimos locais, melhorando os resultados da avaliação.

Além destes, novas estruturas podem ser propostas para a conversão do método de estudo de processos PSCM/ACO da forma supervisionada para a forma assistida. Na forma como a estratégia encontra-se concebida, são necessários valores de referência para as variáveis de estado de interesse das respectivas amostras em estudo. Esta característica implica em que um conjunto de amostras deva ser caracterizado previamente, para então ser analisado. A proposta de expansão da técnica para a versão assistida propõe combinar a metodologia de caracterização PSCM/ACO com simulação e ajuste de modelos dinâmicos de processos, como no trabalho de Oliveira *et al.* (2008). Métodos assistidos possuem vantagens com relação a métodos supervisionados, como por exemplo, a não necessidade de caracterizar todas as amostragens com relação às variáveis de interesse.

Neste sistema, propõe-se a realização conjunta entre a estratégia PSCM/ACO, buscando regiões e grupos de elementos espectrais que combinados possuam capacidade preditiva elevada com relação a determinada variável de estado. Esta variável, por usa vez, é simulada através de um modelo proposto para o processo em questão de forma a fornecer seu vetor de valores. A predição apresentada pelo melhor resultado de PSCM/ACO é comparada com a simulação do modelo. A diferença entre estes vetores permite a atualização dos parâmetros do modelo de estados em conjunto com a seleção do grupo de elementos espectrais para caracterização do processo com dados de espectroscopia.

A versão assistida, em comparação à versão supervisionada, necessita apenas de medidas suficientes de dados espectroscópicos e de uma estrutura de modelo em espaço de estados satisfatória do processo em questão. Como resultados, podem ser ajustados os parâmetros do modelo em espaço de estados, simultaneamente com a qualificação e seleção das regiões espectrais significativas associadas a cada variável de estado.

Com relação à aplicação da técnica PSCM/ACO na configuração de janela móvel, como ferramenta para a caracterização e seleção de regiões espectrais, a evolução natural é a implementação do método para a análise de dados bi-dimensionais, como no caso de dados oriundos de Espectroscopia de Fluorescência 2D.

Além de modificações conceituais do método proposto, testes em diferentes processos, com o intuito de caracterizar estes com relação a dados espectroscópicos, podem ser realizados, permitindo um melhor aproveitamento da ferramenta e de dados de espectroscopia.

Referências

Ait Kaddour, A. and B. Cuq (2009). "In line monitoring of wet agglomeration of wheat flour using near infrared spectroscopy." Powder Technology **190**(1–2): 10-18.

Alford, J. S. (2006). "Bioprocess control: Advances and challenges." Computers & Chemical Engineering **30**(10-12): 1464-1475.

Allegrini, F. and A. C. Olivieri (2011). "A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis." Analytica Chimica Acta **699**(1): 18-25.

Alves, J. C. L. and R. J. Poppi (2013). "Biodiesel content determination in diesel fuel blends using near infrared (NIR) spectroscopy and support vector machines (SVM)." Talanta **104**(0): 155-161.

Ammari, F., R. Bendoula, D. Jouan-Rimbaud Bouveresse, D. N. Rutledge and J.-M. Roger (2014). "3D front face solid-phase fluorescence spectroscopy combined with Independent Components Analysis to characterize organic matter in model soils." Talanta **125**(0): 146-152.

Andries, J. P. M., Y. V. Heyden and L. M. C. Buydens (2013). "Predictive-property-ranked variable reduction in partial least squares modelling with final complexity adapted models: Comparison of properties for ranking." Analytica Chimica Acta **760**(0): 34-45.

Aynsley M, Hofland A, Morris AJ, Montague GA and D. M. C. (1993). "Artificial intelligence and the supervision of bioprocesses (real-time knowledge-based systems and neural networks)." Bioprocess Design and Control.

Aynsley, M., A. Hofland, A. J. Morris, G. A. Montague and C. Di Massimo (1993). "Artificial intelligence and the supervision of bioprocesses (real-time knowledge-based systems and neural networks)." Bioprocess Design and Control.

Bag, N., D. H. X. Yap and T. Wohland (2014). "Temperature dependence of diffusion in model and live cell membranes characterized by imaging fluorescence correlation spectroscopy." Biochimica et Biophysica Acta (BBA) - Biomembranes **1838**(3): 802-813.

Banwell, C. N. (1983). Fundamentals of Molecular Spectroscopy. New York, McGraw-Hill.

Beebe, K. R., R. J. Pell and M. B. Seasholtz (1998). Chemometrics: A pratical guide. New York, Wiley & Sons.

Bequette, B. W. (2003). Process control : modeling, design, and simulation. Upper Saddle River, NJ, Prentice Hall PTR.

Boehl, D., D. Solle, B. Hitzmann and T. Scheper (2003). "Chemometric modelling with two-dimensional fluorescence data for Claviceps purpurea bioprocess characterization." Journal of Biotechnology **105**(1-2): 179-188.

Bosque-Sendra, J. M., L. Cuadros-Rodríguez, C. Ruiz-Samblás and A. P. de la Mata (2012). "Combining chromatography and chemometrics for the characterization and authentication of fats and oils from triacylglycerol compositional data—A review." Analytica Chimica Acta **724**(0): 1-11.

Brereton, R. (2007). Chemmometrics for Pattern Recognition. Chichester, John Wiley & Sons.

Brereton, R. G. (2013). Chemometrics and Statistics: Multivariate Classification Techniques. Reference Module in Chemistry, Molecular Sciences and Chemical Engineering, Elsevier.

Bro, R., F. van den Berg, A. Thybo, C. M. Andersen, B. M. Jørgensen and H. Andersen (2002). "Multivariate data analysis as a tool in advanced quality monitoring in the food production chain." Trends in Food Science & Technology **13**(6–7): 235-244.

Brown, S. D. (2013). Transfer of Multivariate Calibration Models. Reference Module in Chemistry, Molecular Sciences and Chemical Engineering, Elsevier.

Camacho, J., J. Picó and A. Ferrer (2010). "Data understanding with PCA: Structural and Variance Information plots." Chemometrics and Intelligent Laboratory Systems **100**(1): 48-56.

Cariou, V., S. Verdun and E. M. Qannari (2014). "Quadratic PLS regression applied to external preference mapping." Food Quality and Preference **32, Part A**(0): 28-34.

Cerny, V. (1985). Journal of Optimization Theory and Applications

Champe, P. C. and R. A. Harvey (2005). Biochemistry. Philadelphia, Lippincott/Williams & Wilkins.

Chi, Q., Z. Fei, Z. Zhao, L. Zhao and J. Liang (2014). "A model predictive control approach with relevant identification in dynamic PLS framework." Control Engineering Practice **22**(0): 181-193.

Christensen, J., L. Norgaard and C. Lindemann (1995). "On-Line fluorescence spectroscopy and chemometrics for qualitative and quantitative analysis: application in the sugar industry." Spectrosc. **5**.

Clementschitsch, F., K. Jürgen, P. Florentina and B. Karl (2005). "Sensor combination and chemometric modelling for improved process monitoring in recombinant E. coli fed-batch cultivations." Journal of Biotechnology **120**(2): 183-196.

Cocchi, M., M. Corbellini, G. Foca, M. Lucisano, M. A. Pagani, L. Tassi and A. Ulrici (2005). "Classification of bread wheat flours in different quality categories by a wavelet-based feature selection/classification algorithm on NIR spectra." Analytica Chimica Acta **544**(1–2): 100-107.

Cooper, J. B. (1999). "Chemometric analysis of Raman spectroscopic data for process control applications." Chemometrics and Intelligent Laboratory Systems **46**(2): 231-247.

Croney, J. C., D. M. Jameson and R. P. Learmonth (2001). "Fluorescence spectroscopy in biochemistry: teaching basic principles with visual demonstrations." Biochemistry and Molecular Biology Education **29**(2): 60-65.

Danuta, F. (1988). "The Jablonski diagram." Journal of Photochemistry and Photobiology B: Biology **2**(3): 399.

Dayal, B. S. and J. F. MacGregor (1997). "Improved PLS algorithms." J. Chemom. **11**: 73 - 85.

Deneubourg, J. L., S. Aron, S. Goss, J. M. Pasteels and G. Duerinck (1986). "Random behaviour, amplification processes and number of participants: How they contribute to the foraging properties of ants." Physica D: Nonlinear Phenomena **22**(1–3): 176-186.

Diehl, F. C., A. R. Secchi, L. P. Lusa, L. A. R. Muniz and L. G. S. Longhi (2009). "Simulação Operacional de uma Torre de Destilação atmosférica via Aspen Plus e Avaliação de Modelos de Analisadores Virtuais." Controle e Automação.

Dorigo, M., M. Birattari and T. Stuetzle (2006). "Ant colony optimization - Artificial ants as a computational intelligence technique." Ieee Computational Intelligence Magazine **1**(4): 28-39.

Dorigo, M. and C. Blum (2005). "Ant colony optimization theory: A survey." Theoretical Computer Science **344**(2–3): 243-278.

Dorigo, M. and C. Blum (2005). "Ant colony optimization theory: A survey." Theoretical Computer Science **344**(2-3): 243-278.

Dorigo, M., G. Di Caro and T. Stützle (2000). "Ant algorithms." Future Generation Computer Systems **16**(8): v-vii.

Du, Y. P., Y. Z. Liang, J. H. Jiang, R. J. Berry and Y. Ozaki (2004). "Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares." Analytica Chimica Acta **501**(2): 183-191.

Eliasson Lantz, A., P. Jørgensen, E. Poulsen, C. Lindemann and L. Olsson (2006). "Determination of cell mass and polymyxin using multi-wavelength fluorescence." Journal of Biotechnology **121**(4): 544-554.

Everitt, B. S. (2002). *Cambridge Dictionary of Statistics*.

Farrés, M., M. Villagrasa, E. Eljarrat, D. Barceló and R. Tauler (2012). "Chemometric evaluation of different experimental conditions on wheat (Triticum aestivum L.) development using liquid chromatography mass spectrometry (LC–MS) profiles of benzoxazinone derivatives." Analytica Chimica Acta **731**(0): 24-31.

Ferd, W. (1981). "Review of present trends in luminescence research." Journal of Luminescence **24-25, Part 2**(0): 929-936.

Fernandes, H. L., I. M. R. Jr, C. Pasquini and J. J. R. Rohwedder (2008). "Simultaneous determination of methanol and ethanol in gasoline using NIR spectroscopy: Effect of gasoline composition." Talanta **75**(3): 804-810.

Ferraty, F. and Y. Romain (2011). The Oxford handbook of functional data analysis. Oxford ; New York, Oxford University Press.

Filgueiras, P. R., J. C. L. Alves and R. J. Poppi (2014). "Quantification of animal fat biodiesel in soybean biodiesel and B20 diesel blends using near infrared spectroscopy and synergy interval support vector regression." Talanta **119**(0): 582-589.

Filzmoser, P. and V. Todorov (2011). "Review of robust multivariate statistical methods in high dimension." Analytica Chimica Acta **705**(1–2): 2-14.

Gaydou, V., J. Kister and N. Dupuy (2011). "Evaluation of multiblock NIR/MIR PLS predictive models to detect adulteration of diesel/biodiesel blends by vegetal oil." Chemometrics and Intelligent Laboratory Systems **106**(2): 190-197.

Ge, X., Y. Kostov and G. Rao (2005). "Low-cost noninvasive optical CO2 sensing system for fermentation and cell culture." Biotechnology and bioengineering **89 (3)**: 329-334.

Geladi, P. (2003). "Chemometrics in spectroscopy. Part 1. Classical chemometrics." Spectrochimica Acta Part B: Atomic Spectroscopy **58**(5): 767-782.

Geladi, P. and B. R. Kowalski (1986). "Partial least-squares regression: a tutorial." Analytica Chimica Acta **185**(0): 1-17.

Geladi, P., B. Sethson, J. Nyström, T. Lillhonga, T. Lestander and J. Burger (2004). "Chemometrics in spectroscopy: Part 2. Examples." Spectrochimica Acta Part B: Atomic Spectroscopy **59**(9): 1347-1357.

Gemperline, P. (2006). Practical guide to chemometrics. Boca Raton, CRC/Taylor & Francis.

Ghasemi-Varnamkhasti, M., S. S. Mohtasebi, M. L. Rodriguez-Mendez, A. A. Gomes, M. C. U. Araújo and R. K. H. Galvão (2012). "Screening analysis of beer ageing using near infrared spectroscopy and the Successive Projections Algorithm for variable selection." Talanta **89**(0): 286-291.

Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning Boston, Kluwer Academic Publishers.

Gomes, A. d. A., R. K. H. Galvão, M. C. U. Araújo, G. Véras and E. C. Silva (2013). "The successive projections algorithm for interval selection in PLS." Microchemical Journal **110**(0): 202-208.

Goss, S., S. Aron, J. L. Deneubourg and J. M. Pasteels (1989). "Self-organized Shortcuts in the Argentine Ant." Naturwissenschaften **76**: 579 - 581.

Guimet, F., J. Ferré, R. Boqué and F. X. Rius (2004). "Application of unfold principal component analysis and parallel factor analysis to the exploratory analysis of olive oils by means of excitation–emission matrix fluorescence spectroscopy." Analytica Chimica Acta **515**(1): 75-85.

Habibi, A., F. Vahabzadeh and M. Zaiat (2013). "Dynamic mathematical models for biodegradation of formaldehyde by Ralstonia eutropha in a batch bioreactor." Journal of Environmental Management **129**(0): 548-554.

Hafuka, A., H. Yoshikawa, K. Yamada, T. Kato, M. Takahashi, S. Okabe and H. Satoh (2014). "Application of fluorescence spectroscopy using a novel fluoroionophore for quantification of zinc in urban runoff." Water Research **54**(0): 12-20.

Hagerdon, A., R. L. Legge and H. Budman (2003). "Evaluation of spectrofluorometry as a tool for estimation in fed-batch fermentations." Bitechnol. Bioeng. **83**: 104–111.

Hantelmann, K., M. Kollecker, D. Hüll, B. Hitzmann and T. Scheper (2006). "Two-dimensional fluorescence spectroscopy: a novel approach for controlling fed-batch cultivations. ." Journal of Biotechnology **121**: 410-417.

Hasegawa, T. (2001). Handbook of Vibrational Spectroscopy. J. Chalmers and P. R. Griffiths. Chichester, UK, Wiley**:** 2293.

He, W., H. Dou, Z. Li, X. Wang, L. Wang, R. Wang and J. Chang (2014). "Investigation of the interaction between five alkaloids and human hemoglobin by fluorescence spectroscopy and molecular modeling." Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy **123**(0): 176-186.

Hemmateenejad, B., R. Miri and M. Elyasi (2012). "A segmented principal component analysis—regression approach to QSAR study of peptides." Journal of Theoretical Biology **305**(0): 37-44.

Hemmateenejad, B., M. Shamsipur, V. Zare-Shahabadi and M. Akhond (2011). "Building optimal regression tree by ant colony system–genetic algorithm: Application to modeling of melting points." Analytica Chimica Acta **704**(1–2): 57-62.

Hitzmann, B., T. Pekeler, C. Lindermann, S. Marose and T. Scheper (1998). Chemometric models for the on-line estimation of bioprocess variables from 2-D fluorescencespectra. 7th International Conference on Computer Applications in Biotechnology. Osaka.

Horvath, J. J., S. A. Glazier and C. J. Spangler (1993). "In Situ Fluorescence cell mass measurements of Saccharomyces cerevisiae using cellular Tryptophan." Biotechnol. Prog **9**: 666-670.

Hougaard, A. B., A. J. Lawaetz and R. H. Ipsen (2013). "Front face fluorescence spectroscopy and multi-way data analysis for characterization of milk pasteurized using instant infusion." LWT - Food Science and Technology **53**(1): 331-337.

Jackson, J. E. (1991). A User's Guide to Principal Components. Chichester, Wiley.

Jiang, J.-H., R. J. Berry, H. W. Siesler and Y. Ozaki (2002). "Wavelength Interval Selection in Multicomponent Spectral Analysis by Moving Window Partial Least-Squares Regression with Applications to Mid-Infrared and Near-Infrared Spectroscopic Data " Anal. Chem. **74**: 3555 - 3565.

Joe Qin, S. (1998). "Recursive PLS algorithms for adaptive data modeling." Computers & Chemical Engineering **22**(4–5): 503-514.

Jolliffe, I. T. (1986). Principal component analysis. New York, Springer-Verlag.

K., R. and P. P. (2009). "Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques." European Journal of Medicinal Chemistry **44**(7): 2913-2922.

Kabbaj, N., Y. Nakkabi and A. Doncescu (2010). "Analytical and knowledge based approaches for a bioprocess supervision." Knowledge-Based Systems **23**(2): 116-124.

Kara, S., F. Anton, D. Solle, M. Neumann, B. Hitzmann, T. Scheper and A. Liese (2010). "Fluorescence spectroscopy as a novel method for on-line analysis of biocatalytic C–C bond formations." Journal of Molecular Catalysis B: Enzymatic **66 (1-2)**: 124-129.

Kara, S., F. Anton, D. Solle, M. Neumann, B. Hitzmann, T. Scheper and A. Liese (Sep 2010). "Fluorescence spectroscopy as a novel method for on-line analysis of biocatalytic C–C bond formations." Journal of Molecular Catalysis B: Enzymatic **66 (1-2)**: 124-129.

Kessler, W. and R. Kessler (2014) "The Unscrambler® A Handy Tool for Doing Chemometrics ".

Kettaneh, N., A. Berglund and S. Wold (2005). "PCA and PLS with very large data sets." Computational Statistics & Data Analysis **48**(1): 69-85.

Killner, M. H. M., J. J. R. Rohwedder and C. Pasquini (2011). "A PLS regression model using NIR spectroscopy for on-line monitoring of the biodiesel production reaction." Fuel **90**(11): 3268-3273.

Kirkpatrick, S., C. D. Gelatt and M. P. Vecchi (1983). "Optimization by simulated annealing." Science.

Kowalski, B. R. (1983). Chemometrics, Mathematics and Statistics in Chemistry. Consenza, Springer.

Krishnan, A., L. J. Williams, A. R. McIntosh and H. Abdi (2011). "Partial Least Squares (PLS) methods for neuroimaging: A tutorial and review." NeuroImage **56**(2): 455-475.

Kumar, N., A. Bansal, G. S. Sarma and R. K. Rawal (2014). "Chemometrics tools used in analytical chemistry: An overview." Talanta(0).

Lacerda, E. (2007). Otimização Nuvem de Partículas (Particle Swarm). Universidade Federal do Rio Grande do Norte.

Lakowicz, J. R. (2006). Principles of fluorescence spectroscopy. New York, Springer.

Land Jr, W. H., F. W., P. W., M. R., H. N., H. J., E. S., Q. X. and Y. T (2011). "Partial Least Squares Applied to Medical Bioinformatics." Procedia Computer Science **6**(0): 273-278.

Leardi, R., M. B. Seasholtz and R. J. Pell (2002). "Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data." Analytica Chimica Acta **461**(2): 189-200.

Li Vigni, M., C. Durante, G. Foca, A. Marchetti, A. Ulrici and M. Cocchi (2009). "Near Infrared Spectroscopy and multivariate analysis methods for monitoring flour performance in an industrial bread-making process." Analytica Chimica Acta **642**(1–2): 69-76.

Lindemann, C., S. Marose, H. O. Nielsen and T. Scheper (1998). "2-Dimensional fluorescence spectroscopy for on-line bioprocess monitoring." Sensors and Actuators B: Chemical **51**(1-3): 273-277.

Liu, R. X., J. Kuang, Q. Gong and X. L. Hou (2003). "Principal component regression analysis with spss." Computer Methods and Programs in Biomedicine **71**(2): 141-147.

Mardia, K., J. Kent and J. Bibby (1979). Multivariate Analysis, Academic Press.

Marose, S., C. Lindemann and T. Scheper (1998). "Two-dimensional fluorescence spectroscopy: A new tool for on-line bioprocess monitoring." Biotechnology Progress **14**(1): 63-74.

Masiero, S. S., J. O. Trierweiler, M. Farenzena, M. Escobar, L. F. Trierweiler and C. Ranzan (2013). "Evaluation of wavelength selection methods for 2D fluorescence spectra applied to bioprocesses characterization." Brazilian Journal of Chemical Engineering **30**.

McLeod, G., K. Clelland, H. Tapp, E. K. Kemsley, R. H. Wilson, G. Poulter, D. Coombs and C. J. Hewitt (2009). "A comparison of variate pre-selection methods for use in partial least squares regression: A case study on NIR spectroscopy applied to monitoring beer fermentation." Journal of Food Engineering **90**(2): 300-307.

Mello, P. A. and J. C. C. S. Pinto (2008). Introdução à Modelagem Matemática e Dinâmica Não-Linear de Processos Químicos. Rio de Janeiro, *Escola Piloto Virtual Giuliano Massaran*.

Menezes, J. C. (2011). Process Analytical Technology in Bioprocess Development and Manufacturing. Comprehensive Biotechnology (Second Edition). M.-Y. Editor-in-Chief: Murray. Burlington, Academic Press**:** 501-509.

Mulchandani, A. and A. S. Bassi (1995). "Principles and applications of biosensors for bioprocess monitoring and control." Crit. Rev. Biotechnol **1**: 105-124.

Mulchandani, A. and J. H. T. Luong (1989). "Microbial inhibition kinetics revisited." Enzyme and Microbial Technology **11**(2): 66-73.

Mullen, R. J., D. Monekosso, S. Barman and P. Remagnino (2009). "A review of ant algorithms." Expert Systems with Applications **36**(6): 9608-9617.

Mussatto, S. I., G. Dragone, P. M. R. Guimarães, J. P. A. Silva, L. M. Carneiro, I. C. Roberto, A. A. Vicente, L. Domingues and J. A. Teixeira (2010). "Technological trends, global market, and challenges of bio-ethanol production." Biotechnology Advances **28**: 817–830.

Nascimento, R. S., R. E. S. Froes, N. O. C. e Silva, R. L. P. Naveira, D. B. C. Mendes, W. B. Neto and J. B. B. Silva (2010). "Comparison between ordinary least squares regression and weighted least squares regression in the calibration of metals present in human milk determined by ICP-OES." Talanta **80**(3): 1102-1109.

Ni, W., S. D. Brown and R. Man "A localized adaptive soft sensor for dynamic system modeling." Chemical Engineering Science(0).

Oliveira, F. R. P., K. Goldberg, A. Liese and B. Hitzmann (2008). "Chemometric modelling for process analyzers using just a single calibration sample." Chemometrics and Intelligent Laboratory Systems **94**(2): 118-122.

Olivieri, A. C., H. C. Goicoechea and F. A. Iñón (2004). "MVC1: an integrated MatLab toolbox for first-order multivariate calibration." Chemometrics and Intelligent Laboratory Systems **73**(2): 189-197.

Omary, M. A. and H. H. Patterson (1999). Luminescence, Theory. Encyclopedia of Spectroscopy and Spectrometry (Second Edition). L. Editor-in-Chief: John. Oxford, Academic Press**:** 1372-1391.

Omrani, H., A. E. Dudelzak, B. P. Hollebone and H.-P. Loock (2014). "Assessment of the oxidative stability of lubricant oil using fiber-coupled fluorescence excitation–emission matrix spectroscopy." Analytica Chimica Acta **811**(0): 1-12.

Otsuka, M. (2004). "Comparative particle size determination of phenacetin bulk powder by using Kubelka–Munk theory and principal component regression analysis based on near-infrared spectroscopy." Powder Technology **141**(3): 244-250.

Pasquini, C. (2002). Espectroscopia no Infravermelho Proximo (NIR). Salvador, UFBA.

Pattison, R. N., J. Swamy, B. Mendenhall, C. Hwang and B. T. Frohlich (2000). "Measurement and control of dissolved carbon dioxide in mammalian cell culture processes using an in situ fiber optic chemical sensor." Biotechnology progress **16 (5)**: 769-774.

Pereira, J. P. G. (2007). Heurísticas computacionais aplicadas à otimização estrutural de treliças bidimensionais Centro Federal de Educação Tecnológica de Minas Gerais.

Pratap R, P. (2003). "Oscillatory metabolism of Saccharomyces cerevisiae: an overview of mechanisms and models." Biotechnology Advances **21**(3): 183-192.

Pratap, R. P. (2003). "Oscillatory metabolism of Saccharomyces cerevisiae: an overview of mechanisms and models." Biotechnology Advances **21**(3): 183-192.

Química, A.-A. B. d. I. (2013). A Indústria Química Brasileira.

Ramos, L. S., K. R. Beebe, W. P. Carey, E. M. Sanchez, B. C. Erickson, B. E. Wilson, L. E. Wangen and B. R. Kowalski (1986). "Chemometrics." Anal. Chem. **58**: 294 - 315.

Ranzan, C. (2010). Fermentação Contínua de Zymomonas mobilis: Modelagem, Ajuste de Parâmetros e Inferências a Partir do Consumo de Hidróxido de Sódio. Master, Universidade Federal do Rio Grande do Sul.

Ranzan, C., L. Ranzan, L. F. Trierweiler, B. Hitzmann and J. O. Trierweiller (2013). Caracterizacao de Fermentacoes de gllicose com Saccharomyces cerevisiae utilizando Espectroscopia Fluorescente 2D e Ant Colony Optimization. Foz do Iguacu, PR, Brasil, 19 Simpósio Nacional de Bioprocessos.

Ranzan, C., A. Strohm, L. Ranzan, L. F. Trierweiler, B. Hitzmann and J. O. Trierweiler (2014). "Wheat flour characterization using NIR and spectral filter based on Ant Colony Optimization." Chemometrics and Intelligent Laboratory Systems **132**(0): 133-140.

Ranzan, C., L. Trierweiler, B. Hitzmann and J. O. Trierweiler (2011). Validação com Dados de Espectroscopia Fluorescente 2D de Modelo Dinâmico para Fermentações Batelada de Saccharomyces cerevisiae. Caxias do Sul, RS, Brasil, 18 Simpósio Nacional de Bioprocessos.

Ranzan, C., L. Trierweiler and J. O. Trierweiler (2011). Modelagem quimiométrica para para caracterização de bioprocessos através de espectroscopia fluorescente-2D. Caxias do Sul, RS, Brasil, 18 Simpósio Nacional de Bioprocessos 2011.

Ranzan, C., L. F. Trierweiler, B. Hitzmann and J. O. Trierweiler (2012). Fluorescence Spectroscopy as a Tool for Ethanol Fermentation On-line Monitoring. International Symposium on Advanced Control of Chemical Processes. Singapore.

Reddy, T. A. (2011). Applied Data Analysis and Modeling for Energy Engineers and Scientists. Arizona, Springer.

Rhee, J. I. and T.-H. Kang (2007). "On-line process monitoring and chemometric modeling with 2D fluorescence spectra obtained in recombinant E. coli fermentations." Process Biochemistry **42**(7): 1124-1134.

Rhee, J. I., A. Ritzka and T. Scheper (2004). "On-Line Monitoring and Control of Substrate Concentrations in Biological Processos by Flowinjection Analysis System." Biotechnol. Bioproc. **9**: 156–165.

Riley, M. R., C. D. Okeson and B. L. Frazier (1999). "Rapid calibration of near-infrared spectroscopic measurements of mammalian cell cultivations." Biotechnology progress **15 (6)**: 1133-1141.

Roberts, M. S., Y. Dancik, T. W. Prow, C. A. Thorling, L. L. Lin, J. E. Grice, T. A. Robertson, K. König and W. Becker (2011). "Non-invasive imaging of skin physiology and percutaneous penetration using fluorescence spectral and lifetime imaging with multiphoton and confocal microscopy." European Journal of Pharmaceutics and Biopharmaceutics **77**(3): 469-488.

Roy, K. and P. Pratim Roy (2009). "Comparative chemometric modeling of cytochrome 3A4 inhibitory activity of structurally diverse compounds using stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques." European Journal of Medicinal Chemistry **44**(7): 2913-2922.

Sato, H., M. Kiguchi, F. Kawaguchi and A. Maki (2004). "Practicality of wavelength selection to improve signal-to-noise ratio in near-infrared spectroscopy." NeuroImage **21**(4): 1554-1562.

Scheper, T., B. Hitzmann, E. Stärk, R. Ulber, R. Faurie, P. Sosnitza and K. F. Reardon (1999). "Bioanalytics: detailed insight into bioprocesses." Analytica Chimica Acta **400**: 121-134.

Schügerl, K. (2001). "Progress in monitoring, modeling and control of bioprocesses during the last 20 years." Journal of Biotechnology **85**(2): 149-173.

Science., B. R. S. o. t. C. o. F. (1995). Biotechnology for the 21st Century: New Horizons. Washington, DC., *National Science and Technology Council*, **:** 41-42.

Settle, F. (1997). Handbook of Instrumental Techniques for Analytical Chemistry. Upper Saddle River, Prentice-Hall.

Shamsipur, M., V. Zare-Shahabadi, B. Hemmateenejad and M. Akhond (2006). "Ant colony optimisation: a powerful tool for wavelength selection." J. Chemometrics **20**: 12.

Shamsipur, M., V. Zare-Shahabadi, B. Hemmateenejad and M. Akhond (2009). "An efficient variable selection method based on the use of external memory in ant colony optimization. Application to QSAR/QSPR studies." Analytica Chimica Acta **646**(1–2): 39-46.

Siesler, H. W., Y. Ozaki, S. Kawata and H. M. Heise (2002). Near-Infrared Spectroscopy, Wiley-VCH.

Silva, M. M. (2011). Otimização de Estruturas Reticuladas Incluindo Não-Linearidade Geométrica, Universidade Federal de Juiz de Fora

Skibsted, E., C. Lindemann, C. Roca and L. Olsson (2001). "On-line bioprocess monitoring with a multi-wavelength fluorescence sensor using multivariate calibration." Journal of Biotechnology **88**(1): 47-57.

Skoog, D. A., F. J. Holler and S. R. Crouch (2007). Principles of instrumental analysis. Belmont, CA, Thomson Brooks/Cole.

Socha, K. and M. Dorigo (2008). "Ant colony optimization for continuous domains." European Journal of Operational Research **185**(3): 1155-1173.

Solle, D., D. Geissler, E. Stärk, T. Scheper and B. Hitzmann (2003). "Chemometric Modelling based on 2D-Fluorescence Spectra without a Calibration Measurement." Bioinformatics **19**: 173-177.

Sorouraddin, M.-H., M.-R. Rashidi, E. Ghorbani-Kalhor and K. Asadpour-Zeynali (2005). "Simultaneous spectrofluorimetric and spectrophotometric determination of melatonin and pyridoxine in pharmaceutical preparations by multivariate calibration methods." Il Farmaco **60**(5): 451-458.

Sotomayor, M. D. P. T., I. L. T. Dias, M. R. V. Lanza, A. B. Moreira and L. T. Kubota (2008). "Aplicação e avanços da espectroscopia de luminescência em análises farmacêuticas." Química Nova **31**(7): 1755-1774.

Sratthaphut, L. and N. Ruangwises (2012). "Genetic Algorithms-Based Approach for Wavelength Selection in Spectrophotometric Determination of Vitamin B12 in Pharmaceutical Tablets by Partial Least-Squares." Procedia Engineering **32**(0): 225-231.

Strohm, A. (2012). Untersuchung von NIR - und Fluoreszenzspektren zur Charakterisierung von Mehlen, Universität Hohenheim.

Stärk, E., B. Hitzmann, K. Schügerl, T. Scheper, C. Fuchs, D. Köster and H. Märkl (2002). "In-Situ-fluorescence-probes: A useful tool for non-invasive bioprocess monitoring." Advances in biochemical engineering Biotechnology **74**: 21 - 38.

Sun, D.-W. (2008). Infrared Spectroscopy for Food Quality Analysis and Control, Elsevier.

Surribas, A., D. Geissler, A. Gierse, T. Scheper, B. Hitzmann, J. L. Montesinos and F. Valero (2006). "State variables monitoring by in situ multi-wavelength fluorescence spectroscopy in heterologous protein production by Pichia pastoris." Journal of Biotechnology **124**(2): 412-419.

Tartakovsky, B., L. A. Lishman and R. L. Legge (1996). "Application of multi-wavelength fluorometry for monitoring wastewater treatment process dynamics." Water Research **30**(12): 2941-2948.

Technology, I.-I. A. f. C. S. a. (1986). Procedure for near infrared (NIR) reflectance analysis of ground wheat and milled wheat products. **202**.

Technology, I.-I. A. f. C. S. a. (1995). Determination of Protein by Near Infrared Reflectance (NIR) Spectroscopy. **159**.

Tellinghuisen, J. (2010). "Least-squares analysis of data with uncertainty in x and y: A Monte Carlo methods comparison." Chemometrics and Intelligent Laboratory Systems **103**(2): 160-169.

Tipler, P. A. and R. A. Llewellyn (2008). Modern physics. New York, NY, W.H. Freeman.

Upton, G. J. G., I. Cook and Oxford University Press. (2008). A dictionary of statistics. Oxford paperback reference. Oxford, Oxford University Press,**:** 1 online resource.

Vaidyanathan, S., S. A. Arnold, L. Matheson, P. Mohan, B. McNeil and L. M. Harvey (2001). "Assessment of near-infrared spectral information for rapid monitoring of bioprocess quality." Biotechnology and bioengineering **74 (5)**: 376-388.

Valencia, S., J. M. Marín, G. Restrepo and F. H. Frimmel (2014). "Evaluation of natural organic matter changes from Lake Hohloh by three-dimensional excitation–emission matrix fluorescence spectroscopy during TiO2/UV process." Water Research **51**(0): 124-133.

Varmuza, K. and P. Filzmoser (2008). Introduction to multivariate statistical analysis in chemometrics. Boca Raton, Taylor & Francis,**:** 1 online resource (xiii, 321 p. ill.).

Vigni, M. L., C. Baschieri, G. Foca, A. Marchetti, A. Ulrici and M. Cocchi (2011). Chapter 2 - Monitoring Flour Performance in Bread Making. Flour and Breads and their Fortification in Health and Disease Prevention. P. Victor, R. W. Ronald Watson and Vinood PatelA2 - Victor Preedy and P. Vinood. San Diego, Academic Press**:** 15-25.

Wang, S. (2010). "Infrared Spectroscopy for Food Quality Analysis and Control." Trends in Food Science & Technology **21**(1): 52.

Wang, X., Q. Yan, P. Chu, Y. Luo, Z. Zhang, S. Wu, L. Wang and Q. Zhang (2011). "Analysis on fluorescence of dual excitable Eu(TTA)3DPBT in toluene solution and PMMA." Journal of Luminescence **131**(8): 1719-1723.

Warth, B., G. Rajkai and C.-F. Mandenius (2010). "Evaluation of software sensors for on-line estimation of culture conditions in an Escherichia coli cultivation expressing a recombinant protein." Journal of Biotechnology **147**(1): 37-45.

Wehrens, R. (2011). Chemometrics With R: Multivariate Data Analysis in the Natural Sciences and Life Sciences, Springer.

Whitford, W. and C. Julien (2007). "Analitical Technology and PAT." BioProcess International: 32-41.

Wolf, G., J. S. Almeida, C. Pinheiro, V. Correia, C. Rodrigues, M. *A.* Reis and J. G. Crespo (2001). "Two-dimensional fluorometry coupled with artificial neural networks: a novel method for on-line monitoring of complex biological processes." Biotechnol. Bioeng **72**: 297–306.

Wong, W. C., C. C. Chan, P. Hu, J. R. Chan, Y. T. Low, X. Dong and K. C. Leong (2014). "Miniature pH optical fiber sensor based on waist-enlarged bitaper and mode excitation." Sensors and Actuators B: Chemical **191**(0): 579-585.

Xu, L. and I. **Schechter** (1996). "Wavelength selection for simultaneous spectroscopic analysis. Experimental and theoretical study." **Anal Chem 68**: 2392–2400.

Yamuna, R. K. and R. V. S. Ramachandra (1999). "Control of fermenters – a review." Bioprocess Engineering **21**: 77-88.

Zang, Z., J. M. Scharer and M. Moo-Young (1997). "Mathematical model for aerobic culture of a recombinant yeast." Bioprocess Engineering **17**: 235–240.

Zhang, Y., A. M. Zamamiri, M. A. Henson and M. A. Hjortsø (2002). "Cell population models for bifurcation analysis and nonlinear control of continuous yeast bioreactors." Journal of Process Control **12**(6): 721-734.

Župerl, Š., S. Fornasaro, M. Novič and S. Passamonti (2011). "Experimental determination and prediction of bilitranslocase transport activity." Analytica Chimica Acta **705**(1–2): 322-333.