SAMUEL ANDERSON ALVES DE SOUSA

NOVAS METODOLOGIAS PARA A ANÁLISE DE DADOS EM CIÊNCIAS ÔMICAS E PARA O CONTROLE DE QUALIDADE DE AMOSTRAS DE BIODIESEL-DIESEL

CAMPINAS

2013

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE QUÍMICA

SAMUEL ANDERSON ALVES DE SOUSA

NOVAS METODOLOGIAS PARA A ANÁLISE DE DADOS EM CIÊNCIAS ÔMICAS E PARA O CONTROLE DE QUALIDADE DE AMOSTRAS DE BIODIESEL-DIESEL

ORIENTADORA: PROFA. DRA. MÁRCIA MIGUEL CASTRO FERREIRA
CO-ORIENTADOR: PROF. DR. ALVICLÉR MAGALHÃES

TESE DE DOUTORADO APRESENTADA AO
INSTITUTO DE QUÍMICA DA UNICAMP PARA
OBTENÇÃO DO TÍTULO DE DOUTOR EM CIÊNCIAS.

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA POR SAMUEL ANDERSON ALVES DE SOUSA, E ORIENTADA PELA PROFA. DRA. MÁRCIA MIGUEL CASTRO FERREIRA.

_____

Assinatura do Co-orientador

CAMPINAS
2013

FICHA CATALOGRÁFICA ELABORADA POR DANIELLE DANTAS DE SOUSA - CRB8/6490 - BIBLIOTECA DO INSTITUTO DE QUÍMICA DA UNICAMP

So85n

Sousa, Samuel Anderson Alves de (1983-).
Novas metodologias para a análise de dados em ciências ômicas e para o controle de qualidade de amostras de biodiesel-diesel / Samuel Anderson Alves de Sousa. – Campinas, SP: [s.n.], 2013.

Orientador: Márcia Miguel Castro Ferreira.
Coorientador: Alviclér Magalhães.

Tese (doutorado) - Universidade Estadual de Campinas, Instituto de Química.

1. Biodiesel. 2. Ressonância magnética nuclear. 3. Bucketing otimizado. 4. Ciências ômicas. 5. Escalamento de diferenças individuais multinível. I. Ferreira, Márcia Miguel Castro. II. Magalhães, Alviclér. III. Universidade Estadual de Campinas. Instituto de Química. IV. Título.

Informações para Biblioteca Digital

Título em inglês: New methodologies for data analysis in omics sciences and for quality control of biodiesel-diesel samples

Palavras-chave em inglês:
Biodiesel
Nuclear magnetic resonance
Optimized bucketing
Omic sciences
Multilevel individual differences scaling

Área de concentração: Físico-Química

Titulação: Doutor em Ciências

Banca examinadora:
Alviclér Magalhães [Coorientador]
Norberto Peporine Lopes
Edenir Rodrigues Pereira Filho
Jarbas José Rodrigues Rohwedder
Roy Edward Bruns

Data de defesa: 25/09/2013

Programa de pós-graduação: Química

À minha mãe Izabel,
ao meu pai João Alberto (*in memoriam*),
ao meu irmão Fred e aos meus sobrinhos
Jhennifer, João Alberto Neto e Afonso.

Agradecimentos

Meus sinceros agradecimentos:

Em primeiro lugar a Deus que na sua infinita bondade tem guiado meus passos e sido o meu pastor.

À Profa. Márcia Miguel, o apoio, orientação e amizade.

Ao Prof. Alviclér Magalhães, a disponibilidade para discussões e a valiosa co-orientação.

Aos Profs. Johan A. Westerhuis, Age K. Smilde e Gooitzen Zwanenburg do *Biosystems Data Analysis Group* do *Swammerdam Institute for Life Sciences* da *University of Amsterdam*, em especial ao primeiro, a orientação durante o estágio de doutorado sanduíche. Deste mesmo grupo, manifesto agradecimentos aos amigos doutorandos e pós-doutorandos, em especial à amiga Dicle Hasdemir.

Ao Prof. José Dias de Souza Filho (ICEx – UFMG), a ajuda com os experimentos de RMN.

A todos os amigos do Laboratório de Quimiometria Teórica e Aplicada (LQTA) do Instituto de Química da Unicamp, em especial ao João Paulo e ao Clecio Filho, pois mais que amigos, são verdadeiros irmãos.

Aos amigos: Letícia, Adriano (Sol), Sérgio, Thiaguim, Flamys, Lucas, Irlene, Janaína, Michel, Edmundo, Nêgo Chico (Chicão), Lenilson, Olímpio, Juliana (do JP), Eduardo, Herbert e família, Reginaldo e família, Márcia, Alessandra, Elidiane, Lilia, Euzébio, Eva, Georgiana e Socorro, o convívio, amizade, companheirismo e todas as outras ótimas coisas que me proporcionaram, no sentido de amenizar as dificuldades da vida longe de casa.

Ao IQ-UNICAMP, todo o suporte fornecido e à CPG-IQ-UNICAMP, em especial aos funcionários Bel e Miguel, o apoio ao longo destes anos.

À Capes, a bolsa de estudos de pós-graduação (Portaria nº 87 - DOU 28/09/2006) e ao CNPq, a bolsa de doutorado sanduíche (Programa Ciência sem Fronteiras).

CURRICULUM VITAE

Dados pessoais:

Nome: Samuel Anderson Alves de Sousa

Filiação: João Alberto de Sousa e Izabel Cristina Alves de Sousa

Nascimento: 10/08/1983 - Teresina/PI - Brasil

Formação acadêmica/titulação:

- **03/2008 – 08/2013: Doutorado em Química** - Universidade Estadual de Campinas. Bolsista da Capes. Orientadora: Profa. Dra. Márcia Miguel Castro Ferreira.
- **02/2006 – 02/2008: Mestrado em Química** - Universidade Federal do Piauí. Orientador: Prof. Dr. José Machado Moita Neto.
- **03/2001 - 06/2005: Graduação em Bacharelado em Química com Atribuições Tecnológicas** - Universidade Federal do Piauí.

Experiência internacional:

Estágio de doutorado sanduíche no *Biosystems Data Analysis Group* do *Swammerdam Institute for Life Sciences* da *University of Amsterdam* – Holanda, sob orientação do Prof. Johan A. Westerhuis, no período de Setembro a Dezembro de 2011.

Atuação profissional:

- **Universidade Estadual de Campinas**:
  Programa de estágio docente, PED B, Período: 08/2008 – 11/2008.
  Programa de estágio docente, PED B, Período: 03/2009 – 06/2009.
- **Instituto Dom Barreto** – IDB – Teresina-PI:
  Docente ensino médio, Período: 02/2005 – 12/2007.
- **Colégio Sagrado Coração de Jesus** – CSCJ – Teresina-PI:
  Docente ensino médio, Período: 02/2005 – 12/2007.

Revisor de periódico:

2012 – Atual: Periódico: Journal of Chemometrics (Print).

Prêmios e títulos:

- **2006:** Melhor desempenho no Exame Nacional de Desempenho dos Estudantes (Enade-2005) como concluinte na área de Química, INEP-MEC (Agraciado com bolsa de pós-graduação da Capes - Portaria nº 87 - DOU 28/09/2006).

- **2005:** Láurea Universitária - Honra ao mérito enriquecedor do meio acadêmico, à Universidade Federal do Piauí e à sociedade piauiense, Universidade Federal do Piauí.
- **2003:** Menção Honrosa na XVI Jornada Brasileira de Iniciação Científica em Química com o trabalho "Constituintes químicos da própolis produzida na cidade de Pio IX - PI", Associação Brasileira de Química.

Produção bibliográfica:

Artigos completos publicados em periódicos:

- **SOUSA, SAA**, Magalhães, A, Ferreira, MMC. Optimized bucketing for NMR spectra: Three case studies. *Chemometrics and Intelligent Laboratory Systems*, 122, 93-102, 2013.
- Muniz Filho, RCD, **SOUSA, SAA,** Pereira, FS, Ferreira, MMC. Theoretical Study of Acid-Catalyzed Hydrolysis of Epoxides. *The Journal of Physical Chemistry A*, 114, 5187-5194, 2010.
- **SOUSA, SAA**, Citó, AMGL, Moita Neto, JM. Análise por espectrometria de massas do óleo de mamona produzido em Teresina-PI. *Química no Brasil*, 1, 53-56, 2007.
- Pires, JEP, Fernandes, RM, Fernandes, MZLCM, Viana, GEN, Dourado, JCL, **SOUSA, SAA**. Determinação da concentração inibitória média (CI50) do extrato aquoso de Simarouba versicolor, St. Hill sobre a ovipostura do carrapato bovino (Boophilus microplus, Canestrine, 1887). *Revista Brasileira de Plantas Medicinais*, 9, 23-26, 2007.
- **SOUSA, SAA,** Citó, AMGL, Lopes, JAD. Constituintes do óleo essencial da própolis produzida na cidade de Pio IX - PI. *Revista Brasileira de Plantas Medicinais*, 8, 1-3, 2006.
- Luz Júnior, GE, **SOUSA, SAA,** Moita, GC, Moita Neto, JM. Química geral experimental: uma nova abordagem didática. *Química Nova*, 27, 1, 164-168, 2004.
- Citó, AMGL, Souza, AA, Lopes, JAD, Chaves, MH, Costa, FB, **SOUSA, SAA,** Amaral MPM. Resina de Protium heptaphyllum March (Burceraceae): Composição Química do óleo essencial e avaliação citotóxica frente a Artemia sallina Leach. *Anais da Associação Brasileira de Química*, 52, 2, 74-76, 2003.

Resumos publicados em anais de congressos:

- **SOUSA, SAA,** Muniz Filho, RCD, Ferreira, MMC. Energética envolvida no rearranjo intramolecular indólico na biossíntese da violaceína: Um estudo teórico. *Simpósio Brasileiro de Química Teórica*, 2009, Poços de Caldas - MG.

- Souza, AM, Godoy, LAF, Breitkreitz, MC, **SOUSA, SAA,** Rothschild, L, Poppi, RJ. Alinhamento de cromatogramas em CLAE para análise multivariada. *33ª Reunião Anual da Sociedade Brasileira de Química*, 2010, Águas de Lindóia - SP.
- Martins, JPA, **SOUSA, SAA,** Ferreira, MMC, Moita Neto, JM. Estudo teórico da atividade de tiossemicarbazonas contra Salmonella typhimurium. *Simpósio Brasileiro de Química Teórica*, 2007, Poços de Caldas - MG.
- **SOUSA, SAA,** Citó, AMGL, Lopes, JAD. Estudo da própolis produzida na cidade de Esperantina-PI. *XIII Seminário de iniciação científica da UFPI*, 2004, Teresina - PI.
- **SOUSA, SAA**, Citó, AMGL, Lopes, JAD, Chaves, MH. Constituintes químicos da própolis produzida na cidade de Pio IX - PI. *XLIII Congresso Brasileiro de Química*, 2003, Ouro Preto - MG.
- **SOUSA, SAA**, Citó, AMGL, Lopes, JAD. Constituintes químicos do óleo essencial da própolis produzida na cidade de Pio IX - PI. *II Simpósio Brasileiro de óleos essenciais*, 2003, Campinas - SP.
- **SOUSA, SAA**, Citó, AMGL, Lopes, JAD, Chaves, MH, Souza, CML. Estudo da própolis produzida na cidade de Pio IX-PI. *XII Seminário de iniciação científica da UFPI*, 2003, Teresina - PI.
- **SOUSA, SAA**, Citó, AMGL, Moita Neto, JM. Análise do óleo de mamona por espectrometria de massas. *XLII Congresso Brasileiro de Química*, 2002, Rio de Janeiro - RJ.
- Luz Júnior, GE, Moita, GC, Moita Neto, JM, **SOUSA, SAA**. A Primeira Disciplina Experimental em Química. XI Encontro Nacional de Ensino de Química, 2002, Recife - PE.

Participações em eventos:

- Workshop de Química Analítica de Processos. 2010. (Seminário).
- Workshop on Recent Trends in Computer Simulations of Biomolecular Systems. 2009. (Seminário).
- XV Simpósio Brasileiro de Química Teórica. 2009. (Simpósio).
- 30ª Reunião Anual da Sociedade Brasileira de Química. 2007. (Congresso).
- XLIII Congresso Brasileiro de Química. 2003. (Congresso).
- XLII Congresso Brasileiro de Química. 2002. (Congresso).

RESUMO

Neste trabalho são apresentadas duas novas metodologias multivariadas. Na primeira, é desenvolvida uma ferramenta denominada *bucketing* otimizado para a correção dos desalinhamentos dos espectros de RMN ^{1}H. A análise de componentes principais em intervalos (iPCA) é utilizada para explorar espectros de RMN ^{1}H e ^{13}C. Para a diminuição de ruído destes últimos é utilizada a análise de componentes principais em múltiplas escalas (MSPCA). Os modelos iPCA são construídos para as classes de amostras, metropolitanas e não metropolitanas, em conjunto e separadas, atuando complementarmente na detecção de amostras não conformes. Neste contexto, os padrões espectrais apontaram amostras, previamente reprovadas pelos parâmetros físico-químicos próprios do campo de biocombustíveis. Adicionalmente, os modelos reprovaram amostras com padrões espectrais distintos, não reprovadas pelos parâmetros citados. De modo geral, o desempenho dos modelos utilizando os espectros de RMN ^{1}H foi satisfatório. Uma exceção foi a detecção de amostras fora da especificação para o teor de biodiesel, onde as distinções nos espectros não permitiram a discriminação de amostras com teores próximo ao limite. Contudo, ao se estender um pouco a faixa sugerida na legislação, os modelos mostraram boa melhoria. Os modelos a partir dos espectros de RMN ^{13}C obtiveram desempenho inferior àqueles citados acima. No segundo estudo é apresentado um novo método denominado escalamento de diferenças individuais multinível (ML-INDSCAL), para analisar a variação intra-individual em dados das ciências ômicas, focando em mudanças nas covariâncias dentro dos grupos experimentais e evidenciando as relações entre as variáveis (BVRs). Como somente a variação intra-individual é usada para revelar as BVRs associadas às mudanças dinâmicas, as interpretações sobre o fenômeno no qual os efeitos se baseiam são melhoradas. Um conjunto de dados simulado é explorado para demonstrar a força do método. O método é também aplicado a um conjunto real de dados de um estudo de expressões genéticas em células expressando a proteína viral R (Vpr) na forma nativa e

com as mutações R80A e F72A/R73A. O procedimento *jack-knife* é explorado na validação dos modelos ML-INDSCAL. O método ML-INDSCAL é o primeiro da literatura que combina a exploração da estrutura multinível do conjunto de dados e a investigação de BVRs e pode fornecer valiosas contribuições no campo de seleção de características.

Palavras-chave: biodiesel; ressonância magnética nuclear; *bucketing* otimizado; ciências ômicas; escalamento de diferenças individuais multinível.

ABSTRACT

In this work, two new multivariate methodologies are presented. In the first approach, a tool named optimized bucketing is developed to correct ^{1}H NMR spectra misalignments. The interval principal component analysis (iPCA) is used in order to explore ^{1}H and ^{13}C NMR spectra. The multiscale principal component analysis (MSPCA) is used for denoising of ^{13}C NMR spectra. The iPCA models are built for two classes of samples, metropolitan and non-metropolitan, together and isolated, complementarily providing out-of-specification samples detections. In this context, the spectral profiles pointed out samples out of specification, in accordance to their previously known physical-chemical parameters from the field of biofuels. Additionally, the models were able to identify samples with distinct spectral profiles, but not rejected by the cited parameters. In general, the iPCA models using ^{1}H NMR spectra presented good performances. An exception involves the detection of out-of-specification samples for biodiesel content, where the distinction on spectra profiles did not allow discrimination of samples when the biodiesel content was close to the allowed limit. Nevertheless, a small extension in the range, adopted by the Brazilian legislation, was enough to produce an improvement. The models from the ^{13}C NMR spectra achieved worse performance than those cited above. In the second study is presented a novel method named multilevel individual differences scaling (ML-INDSCAL) to analyze within-individual variation in omic data, focusing on the changing covariances within groups and evidencing the between variables relationships (BVRs). Since only the within-individual variation is used to reveal the BVRs associated to dynamic changes, the interpretations about the real phenomena underlying the treatment are improved. A simulated data set is explored to demonstrate the strength of the method. Also, the method is applied to a real data set from a study of expression profiles in cell lines expressing wild-type and two mutated (R80A and F72A/R73A strains) Vpr. A version of the jack-knife procedure is explored in order to validate the ML-INDSCAL models.

The ML-INDSCAL is the first method in literature that combines the exploration of the multilevel structure and the BVRs investigation and it can provide valuable insights on the feature selection field.

Keywords: biodiesel; nuclear magnetic resonance; optimized bucketing; omic sciences; multilevel individual differences scaling.

SUMÁRIO

LISTA DE ABREVIATURAS

ALS	*Alternating Least Squares*; Quadrados Mínimos Alternados.
ANOVA	*Analysis of Variance*; Análise de Variância.
ANOVA-PCA ou APCA	ANOVA *Principal Component Analysis*; ANOVA Análise de Componentes Principais.
ANOVA-SCA ou ASCA	ANOVA *Simultaneous Component Analysis*; ANOVA Análise de Componentes Simultânea.
AUROC	*Area Under the Receiver Operator Characteristic*; Área sob a Característica do Operador Receptor.
BMR	*Between Metabolites Relationship*; Relações entre Metabólitos.
BVR	*Between Variables Relationship*; Relações entre Variáveis.
CANDECOMP	*Canonical Decomposition.*
CORCONDIA	*Core Consistency Diagnostic*; Diagnóstico de Consistência do Núcleo.
coshift	*correlation shifting*; deslocamento por correlação.
COW	*Correlation Optimized Warping*; Deformação Otimizada por Correlação.
DTW	*Dynamic Time Warping*; Deformação Temporal Dinâmica.
DWT	*Discrete Wavelet Transform*; Transformada de *Wavelet* Discreta.
FID	*Free Induced Decay*; Decaimento Induzido Livre.
CGEM	Cromatografia Gasosa acoplada à Espectrometria de Massas.
GFHT	*Generalized Fuzzy Hough Transform*; Transformada *Fuzzy Hough* Generalizada.
HIV-1	*Human Immunodeficiency Virus type* 1; Vírus da Imunodeficiência Humana do tipo 1.
icoshift	*interval correlation shifting*; deslocamento por correlação em intervalos.
INDSCAL	*Individual Differences Scaling*; Escalamento de Diferenças Individuais.
iPCA	*interval Principal Component Analysis*; Análise de Componentes Principais em intervalos.
CLEM	Cromatografia Líquida acoplada à Espectrometria de

	Massas.
LTR	*Long Terminal Repeat*; Cadeias Repetidas Terminais (DNA).
MANOVA	*Multivariate* ANOVA; ANOVA Multivariada.
MAPK	*Mitogen-Activated Protein Kinase*; Proteína Quinase Ativada por Mitógenos.
ML-INDSCAL	*Multilevel Individual Differences Scaling*; Escalamento de Diferenças Individuais Multinível.
MLR	*Multiple Linear Regression*; Regressão Linear Múltipla.
MSCA	*Multilevel Simultaneous Component Analysis*; Análise de Componentes Simultânea Multinível.
MSPCA	*Multiscale Principal Component Analysis*; Análise de Componentes Principais em Múltiplas Escalas.
NIPALS	*Nonlinear Iterative Partial Least Squares*; Quadrados Mínimos Parciais Iterativos Não-Lineares.
NMC	*Number of Misclassifications*; Número de erros na Classificação.
OBA	*Optimized Bucketing Algorithm*; Algoritmo de *bucketing* otimizado.
PARAFAC	*Parallel Factor Analysis*; Análise de Fatores Paralelos.
PCA	*Principal Component Analysis*; Análise de Componentes Principais.
PCR	*Principal Component Regression*; Regressão em Componentes Principais.
PLS	*Partial Least Squares*; Quadrados Mínimos Parciais.
PLS-DA	*Partial Least Squares Discriminant Analysis*; Análise de Discriminantes por Quadrados Mínimos Parciais.
PRC	*Principal Response Curves*; Curvas de Respostas Principais.
QMF	*Quadrature Mirror Filter*; Filtro Espelho em Quadratura.
QSAR	*Quantitative Structure Activity Relationship*; Relação Quantitativa entre a Estrutura e a Atividade.
RMN	Ressonância Magnética Nuclear.
RMSEC	*Root Mean Square Error of Calibration*; Raiz Quadrada do Erro Médio da Calibração.
RMSECV	*RMSE of Cross Validation*; RQEM da Validação Cruzada.
RMSEP	*RMSE of Prediction*; RQEM da Predição.

RNA	*Ribonucleic Acid*; Ácido Ribonucleico.
ROC	*Receiver Operator Characteristic*; Característica do Operador Receptor.
SCA	*Simultaneous Component Analysis.*
SCA-P	*Simultaneous Component Analysis with invariant Pattern.*
SMART	*Scaled-to-Maximum, Aligned, and Reduced Trajectories*; Trajetórias Escaladas ao Máximo, Alinhadas e Reduzidas.
SVD	*Singular Value Decomposition*; Decomposição em Valores Singulares.
TUCKER3	*Tucker Decomposition*.
TW	Transformada *Wavelet*.
VL	Variáveis Latentes.
Vpr	*Viral Protein R*; Proteína Viral R.

LISTA DE TABELAS

LISTA DE FIGURAS

Capítulo 1: Base quimiométrica para este trabalho

1.1. Introdução

No presente trabalho de tese são apresentadas novas metodologias para o controle de qualidade de amostras de biodiesel-diesel e para a investigação de dados em ciências ômicas adquiridos segundo um planejamento experimental cruzado ou paralelo. Nos desenvolvimentos destas abordagens são explorados alguns métodos quimiométricos, seja como parte das metodologias propostas ou como ferramentas nas aplicações relacionadas às mesmas. Deste modo, o primeiro capítulo apresenta os aspectos teóricos destes métodos quimiométricos, servindo como base para os demais capítulos. Durante a leitura desta tese, sempre que necessário, o leitor é convidado a retornar às seções deste capítulo para eventuais consultas à teoria.

No segundo capítulo será mostrada uma abordagem denominada *bucketing* otimizado para a correção dos desalinhamentos dos espectros de ressonância magnética nuclear (RMN) de próton, sendo as suas aplicações, citando vantagens e desvantagens, realizadas através de estudos de caso. A análise de componentes principais em intervalos (iPCA), utilizada para explorar espectros de RMN de próton (^{1}H) e carbono (^{13}C), será assunto do terceiro capítulo, onde a metodologia para o controle de qualidade de misturas de biodiesel-diesel será desenvolvida propriamente. Ainda neste assunto, a diminuição de ruído dos dados de RMN ^{13}C será realizada através da análise de componentes principais em múltiplas escalas (MSPCA), que utiliza a transformada *wavelet* e a análise de componentes principais (PCA).

No quarto (e último) capítulo será desenvolvido um novo método, denominado escalamento de diferenças individuais multinível (ML-INDSCAL), para analisar a variação intra-individual em dados das ciências ômicas com medidas repetidas no tempo. A abordagem foca em mudanças nas covariâncias dentro dos grupos

experimentais, com o objetivo de evidenciar as relações entre as variáveis. Dois exemplos serão explorados para demonstrar o método, um conjunto de dados simulado e um conjunto real de dados, disponível na literatura. Uma versão do procedimento *jack-knife* será explorada para a validação dos modelos ML-INDSCAL.

1.2. Quimiometria

Segundo a *International Chemometrics Society* a quimiometria é "a ciência que relaciona as medidas feitas sobre um sistema ou processo químico com o estado do mesmo, através da aplicação de métodos matemáticos ou estatísticos" [1]. De fato, a quimiometria, como a etimologia do nome sugere [2], preocupa-se com dados de origem química, mas atualmente ela tem se tornado um campo da ciência altamente interdisciplinar, aliando diversas áreas, tais como química, estatística, bioinformática, biologia, medicina, engenharia de alimentos, entre outros.

Com o aparecimento de modernos métodos analíticos de alta capacidade e resolução tem ocorrido relevante avanço no que diz respeito à aquisição de grandes quantidades de dados. Neste contexto, é bastante comum a coleta de centenas ou milhares de variáveis por amostra, gerando dados mega ou multivariados, o que por sua vez leva a uma maior exigência aos tratamentos estatísticos, a fim de serem extraídas informações relevantes, razoavelmente validadas, para as inferências científicas [3].

No que diz respeito ao tratamento de conjuntos de dados multivariados, tradicionalmente, são usados os métodos baseados na extração de componentes, e que assim promovem uma redução na dimensionalidade dos dados, evidenciando padrões antes não facilmente percebidos pela inspeção dos conjuntos originais. Tais métodos servem às diversas finalidades em quimiometria, como por exemplo, classificação de amostras, análise exploratória dos dados, construção de modelos de calibração, entre outras.

A análise de componentes principais (PCA, do inglês *Principal Component Analysis*) é de longe o método mais usado para a exploração de dados com as características citadas acima. Detalhes sobre a PCA serão dados a seguir neste capítulo, antes, no entanto, é necessária a definição das notações matemáticas utilizadas ao longo desta tese. As descrições sobre outros métodos explorados nesta tese, além das notações eventualmente necessárias, serão dadas oportunamente neste e nos capítulos seguintes.

1.3. Notações

Tendo em vista o padrão matemático usado em textos na área de quimiometria, utilizam-se as seguintes notações: escalares são definidos em itálico (por exemplo, n); vetores como letras minúsculas em negrito (por exemplo, $\mathbf{y}$) e matrizes como letras maiúsculas em negrito (por exemplo, $\mathbf{X}$). Em alguns casos as matrizes serão escritas explicitamente como $\mathbf{X}$ ($I \times J$) para indicar suas dimensões, com amostras nas I linhas e as variáveis nas J colunas, sendo suas linhas representadas por um vetor $\mathbf{x}_i^{\mathrm{T}}$ com dimensão ($1 \times J$). Os elementos de uma matriz são representados pela letra correspondente à matriz em minúsculo e itálico, sendo as indicações de linha e coluna em subscrito itálico (por exemplo, x_{ij} é um elemento de $\mathbf{X}$ na linha i e coluna j). As colunas de uma matriz são vetores e por isso são representadas pela letra correspondente à matriz em minúsculo e negrito, mas com o número da coluna subscrito itálico (por exemplo, $\mathbf{x}_j$ é a j-ésima coluna de $\mathbf{X}$). Um arranjo de três modos será representado como uma letra maiúscula em negrito e sublinhada (por exemplo, $\underline{\mathbf{X}}$) e os seus elementos representados pela letra correspondente ao arranjo em minúsculo e itálico, sendo as indicações de linha, coluna e fatia em subscrito itálico (por exemplo, x_{ijk} é um elemento de $\underline{\mathbf{X}}$ na linha i, coluna j e fatia k). A matriz identidade será representada por $\mathbf{I}$ com suas dimensões indicadas apropriadamente. Os sobrescritos T e -1 representam as operações transposta e inversa, respectivamente.

1.4. Análise de Componentes Principais

A análise de componentes principais (PCA) é um método de análise exploratória que não faz uso de informações prévias sobre as distribuições das amostras (não supervisionado) e que possibilita a detecção dos padrões essenciais no conjunto de dados. A PCA fundamenta-se na correlação entre as variáveis e assim reduz a dimensionalidade dos dados permitindo que determinadas informações fiquem mais evidentes [4,5].

Do ponto de vista da álgebra linear, a PCA é uma técnica de análise multivariada relacionada à decomposição em autovetores/autovalores e à decomposição em valores singulares (SVD, do inglês *Singular Value Decomposition*). Geometricamente, podemos entender a PCA como uma técnica de projeção, em que uma matriz $\mathbf{X}$ ($I \times J$) é projetada sobre um subespaço de dimensão reduzida, definido por componentes principais [4,5].

Em termos matriciais, PCA é a decomposição da matriz $\mathbf{X}$ ($I \times J$) em duas matrizes de dimensões menores mais uma matriz de resíduos, como mostrado na Equação 1 e na Figura 1.

$$\mathbf{X} = \mathbf{t_1 p_1^T} + \mathbf{t_2 p_2^T} + \cdots + \mathbf{t}_R\mathbf{p}_R^T + \mathbf{E} = \mathbf{TP}^T + \mathbf{E} \qquad \text{Equação 1}$$

Na Equação 1, $\mathbf{T}$ ($I \times R$) é a matriz de escores com colunas $\mathbf{t}_1$, $\mathbf{t}_2$, ..., $\mathbf{t}_R$; $\mathbf{P}$ ($J \times R$) é a matriz de pesos com colunas $\mathbf{p}_1$, $\mathbf{p}_2$, ..., $\mathbf{p}_R$; $\mathbf{E}$ ($I \times J$) é a matriz de resíduos; e R é o número de componentes principais escolhido para descrever os dados originais.

Uma componente principal é definida por um par de pesos e escores, $\mathbf{p}_r$ e $\mathbf{t}_r$, sendo o primeiro responsável pela descrição das relações entre as variáveis e o segundo pela descrição das relações entre as amostras, em cada componente principal. Portanto, similaridades e diferenças entre amostras ou variáveis podem ser vistas através de

gráficos de escores ou pesos, respectivamente, de uma componente principal *versus* outra.

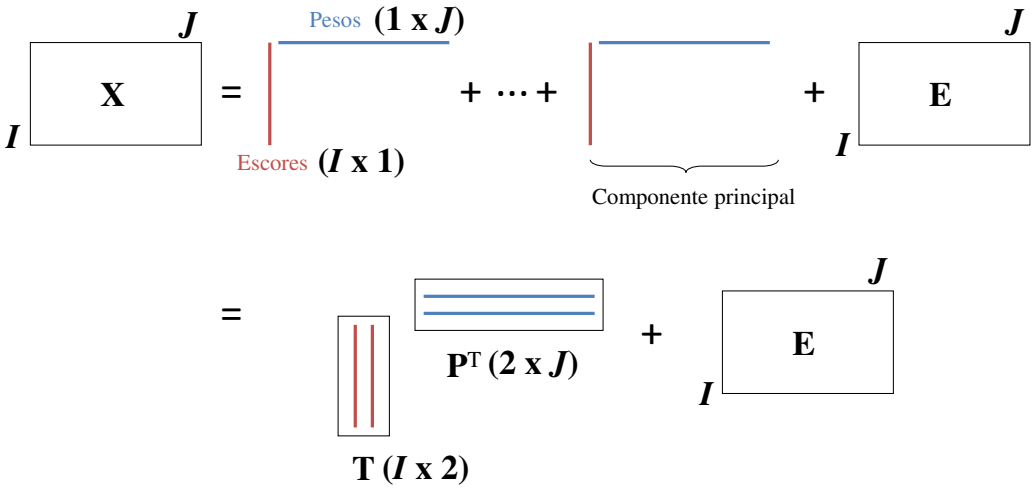

Figura 1. Ilustração para a decomposição de uma matriz **X** ($I \times J$) na PCA onde duas componentes principais ($R = 2$) são escolhidas para descrever os dados.

As componentes principais são obtidas de modo a descrever a máxima variância nos dados e são calculadas em ordem decrescente de importância. Em outras palavras, a primeira componente principal sempre descreve maior variância do que a segunda componente, que por sua vez descreve maior variância que a terceira e assim sucessivamente.

Na decomposição as matrizes obtidas obedecem a certas restrições, a saber, a matriz de escores deve ser ortogonal, ou seja, $\mathbf{T}^T\mathbf{T}$ = matriz diagonal, e a matriz de pesos deve ser ortonormal, ou seja, $\mathbf{P}^T\mathbf{P}$ = matriz identidade, **I**. O que significa que as componentes principais, bem como as informações descritas por elas, são ortogonais entre si [4,5].

O modelo assumido na PCA é um modelo bilinear, pois, como pode ser visto na Equação 1, o lado direito é linear com respeito aos elementos dos vetores de escores e de pesos. Dizemos, portanto, que a PCA é uma decomposição bilinear, pois assume um modelo com esta característica. Em química são comuns dados com comportamento bilinear, portanto, os modelos bilineares são bastante apropriados [4,5].

Geralmente, em química há a aquisição de dados na forma de vetores (sequências de números), sendo cada amostra caracterizada por um destes. Podemos citar como exemplos, os espectros na região do ultravioleta-visível com absorbâncias medidas em certo número comprimentos de onda; cromatogramas registrados como a intensidade de uma resposta (detecção por índice de refração, detecção por quantidade de íons formados num espectrômetro de massas, etc.) pelo tempo; espectros unidimensionais de ressonância magnética de próton e carbono, entre outros. Os métodos bilineares encontram aplicação no tratamento de matrizes que são obtidas a partir da concatenação dos vetores. As matrizes são entidades matemáticas com dois modos, um para as amostras e outro para as variáveis, podendo ser denominadas como entidades de ordem dois ou tensores de segunda ordem. Os vetores são entidades matemáticas de ordem um ou tensores de primeira ordem.

Portanto, observa-se que os dados produzidos em química têm terminologia igual àquela na matemática (mais precisamente na álgebra tensorial), podendo ser definidos ainda os tensores de ordem zero e de ordem três. Neste contexto, se uma medida de um sistema ou processo químico gera um único valor como resposta, ou seja, um escalar, diz-se que este dado é de ordem zero ou um tensor de ordem zero. O arranjo de vários destes dados para diferentes amostras gera um vetor (entidade matemática com um modo). Entre essas medidas temos aquelas feitas com eletrodos íon-seletivos, por exemplo, a medição de pH, e aquelas realizadas com fotômetros que usam filtros simples, entre outras. Neste contexto, o tratamento de dados é feito sobre vetores e assim pode contar com cálculos de média, desvio padrão, mediana e outros [6].

Tensores de terceira ordem ou dados de ordem três são gerados quando na análise, matrizes são obtidas, uma para cada amostra, sendo estas dispostas num arranjo de três modos, algumas vezes chamado de um “cubo de dados”. Neste arranjo, em geral, temos um modo para as amostras e os outros dois para as variáveis (medidas simultaneamente). Isto acontece comumente no uso das chamadas técnicas hifenadas, tais como,

cromatografia gasosa acoplada à espectrometria de massas, espectroscopia de fluorescência multicomprimento de onda, espectroscopia de ressonância magnética nuclear bidimensional, entre outras. No tratamento de dados dos arranjos de três modos técnicas dedicadas são usadas, como a análise de fatores paralelos (PARAFAC, do inglês *Parallel Factor Analysis*) e o método de TUCKER3 [6]. Comentários sobre a PARAFAC serão dados a seguir.

1.5. Análise de Fatores Paralelos

A análise de fatores paralelos (PARAFAC) pode ser entendida como uma generalização da PCA bilinear para a decomposição de arranjos de dados com três ou mais modos. No caso da análise de dados em três modos, a decomposição é feita em tríades ou componentes trilineares, conforme ilustrado na Figura 2.

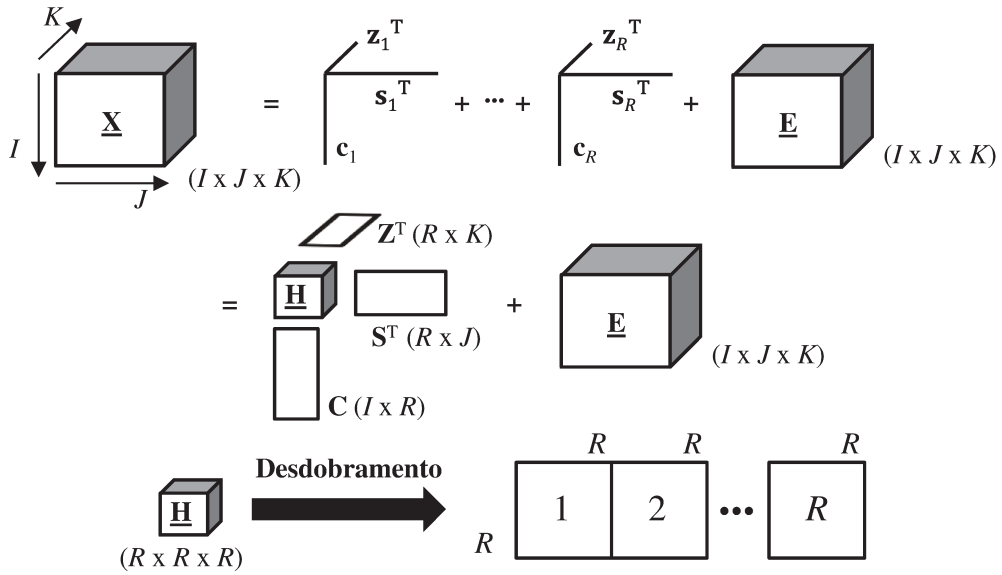

Figura 2. Decomposição de um arranjo de dados **X** com dimensões ($I \times J \times K$) de acordo com um modelo trilinear expresso através de matrizes de pesos, **C** ($I \times R$), **Z** ($K \times R$) e **S** ($J \times R$). **H** é um arranjo central binário de dimensões ($R \times R \times R$), denominado núcleo. Neste caso, R é o número de componentes PARAFAC escolhido para descrever o arranjo original. Também pode ser visto o desdobramento do núcleo **H** em R matrizes com dimensão ($R \times R$) gerando uma matriz **H** ($R \times RR$) desdobrada.

A expressão matemática tradicional (existem outras) para expressar o modelo assumido na PARAFAC necessita da definição prévia de duas operações matriciais, a saber, o produto de Kronecker e o produto de Khatri-Rao com partição em colunas. O produto de Kronecker ou produto tensorial entre duas matrizes quaisquer **A** e **B** com dimensões ($m \times n$) e ($p \times q$) é simbolizado por $\otimes$ e é definido na Equação 2. O resultado é uma matriz de dimensões ($mp \times nq$) [7,8].

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}_{mp \times nq} \qquad \text{Equação 2}$$

O produto de Khatri-Rao com partição em colunas é realizado entre duas matrizes com o mesmo número de colunas e corresponde a um produto de Kronecker em blocos, onde estes blocos são as colunas das matrizes. Nesta tese, o símbolo para este produto de Khatri-Rao é $\odot$ e a operação matemática é mostrada na Equação 3 para duas matrizes **A** e **B** com dimensões ($m \times n$) e ($p \times n$). O resultado é uma matriz de dimensões ($mp \times n$) [7,8].

$$\begin{aligned} \mathbf{A} \odot \mathbf{B} &= [\mathbf{a}_1\ \mathbf{a}_2 \dots \mathbf{a}_n] \odot [\mathbf{b}_1\ \mathbf{b}_2 \dots \mathbf{b}_n] \\ &= [\mathbf{a}_1 \otimes \mathbf{b}_1\ \ \mathbf{a}_2 \otimes \mathbf{b}_2 \dots\ \ \mathbf{a}_n \otimes \mathbf{b}_n]_{mp \times n} \end{aligned} \qquad \text{Equação 3}$$

Portanto, utilizando o produto de Khatri-Rao o modelo PARAFAC para um arranjo de três modos pode ser escrito como mostrado na Equação 4.

$$\mathbf{X}\ (I \times JK) = \mathbf{CH}(\mathbf{Z} \odot \mathbf{S})^{\mathbf{T}} + \mathbf{E}\ (I \times JK) \qquad \text{Equação 4}$$

Na Equação 4 a matriz **X** ($I \times JK$) representa o *unfolding* do mesmo arranjo $\underline{\mathbf{X}}$ ($I \times J \times K$) na Figura 2, que como a própria tradução sugere, corresponde a uma operação onde as fatias do arranjo são colocadas lado a lado "desfolhando ou desdobrando" o mesmo. Observando as dimensões de **X** é possível notar que o desdobramento ocorreu

preservando o modo na direção de I e aumentando a direção J, K vezes. Essa matriz é também chamada de matriz aumentada. As matrizes $\mathbf{C}$ ($I \times R$), $\mathbf{S}$ ($J \times R$) e $\mathbf{Z}$ ($K \times R$) são as mesmas matrizes de pesos na Figura 2 e descrevem as características do arranjo em cada um dos modos. A matriz $\mathbf{E}$ ($I \times JK$) é a matriz de resíduos com as mesmas dimensões de $\mathbf{X}$ [7,9].

Para estimar as matrizes de pesos, diferentes algoritmos podem ser usados, porém o algoritmo baseado em quadrados mínimos alternados (ALS, do inglês *Alternating Least Squares*) é o mais utilizado tendo em vista sua facilidade de implementação e incorporação de restrições, além de uma relativa garantia de convergência, pois ele assegura uma melhoria na resposta a cada iteração [7,10].

O algoritmo ALS procura pelas matrizes de pesos que minimizam a matriz de resíduos, através de uma busca condicionada num processo iterativo que é interrompido quando um determinado critério de convergência é atingido (a menor mudança relativa no ajuste do modelo entre duas iterações consecutivas). Este processo usa a função de ajuste por quadrados mínimos, mostrada na Equação 5, onde $\mathbf{W} = (\mathbf{Z} \odot \mathbf{S})$ [7,9,10].

$$\min_{\mathbf{C},\mathbf{Z},\mathbf{S}} \left\| \mathbf{X}(I \times JK) - \mathbf{C}(\mathbf{Z} \odot \mathbf{S})^{\mathbf{T}} \right\| = \left\| \mathbf{X}(I \times JK) - \mathbf{C}\mathbf{W}^{\mathbf{T}} \right\| \qquad \text{Equação 5}$$

Na Equação 5, o símbolo $\|\cdot\|$ significa a raiz quadrada da soma dos quadrados dos elementos da matriz indicada, ou seja, a norma de Frobenius da matriz. Assim como pode ser visto, a norma de Frobenius da diferença entre a matriz desdobrada $\mathbf{X}$ e o modelo desta matriz, dada pelas matrizes de pesos, é minimizada. Portanto, na busca pelas matrizes de pesos que satisfazem este requisito, uma estimativa para a matriz $\mathbf{C}$ pode ser calculada através de uma estimativa inicial para as matrizes $\mathbf{Z}$ e $\mathbf{S}$ e consequentemente para $\mathbf{W}$, sendo o número de componentes R necessariamente conhecido. A Equação 6 mostra como $\mathbf{C}$ pode ser obtida e devido à simetria do problema a Equação 7 e a Equação 8 mostram como obter as matrizes $\mathbf{Z}$ e $\mathbf{S}$, respectivamente, a cada iteração no algoritmo ALS.

$$\mathbf{C} = \mathbf{X}(I \text{ x } JK)\mathbf{W}(\mathbf{W^T W})^{-1} \qquad \text{Equação 6}$$

$$\mathbf{W} = (\mathbf{C} \odot \mathbf{S}) \text{ e } \mathbf{Z} = \mathbf{X}(K \text{ x } IJ)\mathbf{W}(\mathbf{W^T W})^{-1} \qquad \text{Equação 7}$$

$$\mathbf{W} = (\mathbf{C} \odot \mathbf{Z}) \text{ e } \mathbf{S} = \mathbf{X}(J \text{ x } IK)\mathbf{W}(\mathbf{W^T W})^{-1} \qquad \text{Equação 8}$$

Como observado na Figura 2 o número de colunas das matrizes de pesos (ou o número de tríades em que o arranjo é decomposto) é igual a *R*, o número de componentes do modelo. Desse modo, o valor de *R* deve ser conhecido antes da aplicação do algoritmo ALS. Para situações onde o número de espécies que são responsáveis pelas respostas descritas no arranjo de um sistema químico é conhecido, o valor de *R* será igual ao número de espécies. Em outras situações, que são as mais comuns, o conhecimento exato sobre as espécies do sistema não é possível e assim a estimativa do valor de *R* torna-se uma tarefa não muito fácil. No entanto, é possível estimar *R* baseando-se em alguns critérios como, inspeção visual, variância explicada ou o diagnóstico de consistência do núcleo (CORCONDIA, do inglês *Core Consistency Diagnostic*) [11].

Vale ressaltar que diferentemente da PCA, as tríades ou componentes num modelo PARAFAC não são necessariamente ortogonais entre si e assim a variância explicada pelo modelo não pode ser particionada entre os componentes individuais. Em outras palavras, a inclusão ou retirada de componentes do modelo modifica todas as outras componentes. Como consequência a variância explicada para um modelo PARAFAC somente pode ser relacionada ao modelo como um todo sem partição entre as componentes [10].

A variância explicada é calculada de acordo com a Equação 9, onde no numerador da fração temos os quadrados dos elementos da matriz de resíduos e no numerador o quadrado dos elementos do arranjo $\underline{\mathbf{X}}$.

$$\text{Variância Explicada} = \left(1 - \frac{\sum_{i,j,k}^{I,J,K} e_{ijk}^2}{\sum_{i,j,k}^{I,J,K} x_{ijk}^2}\right) * 100\% \qquad \text{Equação 9}$$

Para o cálculo do CORCONDIA o modelo PARAFAC deve ser considerado como um modelo de TUCKER3 [9] restrito, segundo a Equação 10 [11]:

$$\mathbf{X}(I \times JK) = \mathbf{CG}(R \times RR)(\mathbf{Z} \odot \mathbf{S})^{\mathbf{T}} + \mathbf{E}(I \times JK) \qquad \text{Equação 10}$$

onde a matriz **G** ($R \times RR$) corresponde ao desdobramento do núcleo, a exemplo do que foi mostrado na Figura 2 para **H,** que possui uns na superdiagonal e zero nos demais elementos. Após o cálculo das matrizes de pesos, estas são utilizadas num modelo explicito de TUCKER3, sendo o arranjo central **G** ($R \times RR$) estimado pela minimização da Equação 11:

$$\min_{\mathbf{G}} \left\| \mathbf{X}(I \times JK) - \mathbf{CG}(R \times RR)(\mathbf{Z} \odot \mathbf{S})^{\mathbf{T}} \right\| \qquad \text{Equação 11}$$

Num modelo TUCKER3 [9], o núcleo **G** quantifica a interação entre os modos, através de valores fora da superdiagonal. Portanto, o cálculo de **G** a partir das matrizes de pesos obtidas numa solução por PARAFAC visa verificar o desvio entre os núcleos, **G** e **H**. Deste modo, se o desvio é elevado significa que a inclusão de interações melhora o ajuste do modelo, assim ou as componentes inicias representam alguma variação não sistemática ou na verdade os dados não são apropriados para uma modelagem por PARAFAC e sim por TUCKER3. Em ambos os casos a solução PARAFAC é dita não apropriada, sendo necessário verificar uma solução com um número diferente de componentes. Por outro lado, se a solução PARAFAC, com o número de componentes escolhido, é apropriada, seguramente a trilinearidade é seguida e a inclusão de interações não melhora o ajuste. Neste caso, o núcleo **G** é idealmente igual ao núcleo **H**. Para verificar isto o desvio entre os arranjos pode ser calculado segundo a Equação 12:

$$\text{CORCONDIA} = \left(1 - \frac{\sum_{i,j,k}^{R} (g_{ijk} - h_{ijk})^2}{\sum_{i,j,k}^{R} h_{ijk}^2}\right) * 100\% \qquad \text{Equação 12}$$

Se o modelo PARAFAC é válido então **G** é semelhante a **H**, e o valor de CORCONDIA é próximo de 100%. Se um conjunto de dados não pode ser descrito por um modelo

trilinear ou se um número excessivo de componentes é utilizado **G** será diferente de **H** e o valor de CORCONDIA será próximo de zero ou até mesmo negativo. Para CORCONDIA próximo de 50%, temos um modelo instável.

Diante de tudo que foi exposto, percebe-se que o método PARAFAC encontra-se bastante estudado, tendo suas bases teóricas já bastante consolidadas. De fato, ele tem encontrando várias aplicações na química e em áreas correlatas [9], sendo muito do seu sucesso ligado ao fato de que o modelo assumido é idêntico a muitos modelos de processos físicos que fundamentam vários tipos de dados químicos. Não menos importante, no aspecto matemático, a unicidade das soluções PARAFAC são um grande atrativo em diversas questões e problemas [12-15].

1.6. Modelos de Calibração

A análise de componentes principais e a análise de fatores paralelos são métodos de decomposição utilizados para descrever os dados num subespaço de dimensão reduzida e assim evidenciar informações que servem para análise exploratória das amostras nos dados originais. No entanto, algumas vezes em quimiometria a matriz de dados multivariados é utilizada para estabelecer uma relação quantitativa com uma ou várias respostas (variáveis dependentes), expressando-as através de um modelo matemático que por sua vez é obtido por alguma técnica de regressão. Em geral em química, as variáveis que são modeladas, ou seja, as variáveis dependentes (vetor **y** ou matriz **Y**) são: concentrações de analitos, respostas sensoriais, entre outras, ao passo que a matriz com as variáveis preditoras (matriz **X**), que serão usadas para modelar estas últimas, é proveniente de respostas instrumentais como: espectros de ressonância magnética nuclear [16,17], cromatogramas [18], entre outros. Existem ainda os modelos de calibração na área do estudo das relações quantitativas entre a estrutura química e a atividade biológica (QSAR, do inglês *Quantitative Structure Activity Relationship*) onde

a matriz **X** corresponde a um grupo de variáveis calculadas através de mecânica quântica ou clássica; ou variáveis topológicas e a matriz **Y** (ou vetor **y**) representa um conjunto de atividades biológicas [19].

Em todos os exemplos citados, o problema da regressão, isto é, como modelar uma ou várias variáveis dependentes, numa matriz **Y**, por meio de um conjunto de variáveis preditoras, numa matriz **X**, é solucionado resolvendo a equação $\mathbf{Y} = \mathbf{XB}$, onde **B** é a matriz de regressão. Fazendo uma restrição onde apenas uma variável dependente está presente temos para a mesma equação, $\mathbf{y} = \mathbf{Xb}$, onde **b** é um vetor de regressão contendo os coeficientes que multiplicam as colunas da matriz **X**. A solução da equação é dada por $\mathbf{b} = \mathbf{X}^{+}\mathbf{y}$, onde $\mathbf{X}^{+}$ é a inversa generalizada de Moore-Penrose, ou seja, $\mathbf{X}^{+} = (\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}$ [20].

Se na solução para o problema da regressão apresentado acima a matriz **X** ($I \times J$) é tal que: 1) $I \geq J$, ou seja, existem mais amostras do que variáveis; 2) se todas as colunas são pouco correlacionadas ou aproximadamente linearmente independentes, ou seja, nenhuma delas pode ser escrita como uma combinação linear das outras; 3) e se em **X** há apenas ruído moderado, dizemos que a matriz **X** tem posto completo, no caso, igual a J. Assim, existe um único subespaço J-dimensional gerado pelas colunas de **X**, no qual o vetor **y** pode ser projetado de tal forma que o quadrado da norma-2 da diferença entre ele e a projeção, é o mínimo possível. A projeção corresponde à solução de quadrados mínimos e o método neste caso é chamado de regressão linear múltipla (MLR, do inglês *Multiple Linear Regression*) [21-23].

Por outro lado, se a matriz **X** possui colunas altamente correlacionadas, o que pode ocorrer quando a matriz possui mais variáveis do que amostras ou se as suas colunas são (aproximadamente) linearmente dependentes, dizemos que ela é uma matriz mal condicionada e com posto deficiente. Neste caso, ou são obtidas infinitas soluções de quadrados mínimos (existem infinitos subespaços, nos quais **y** pode ser projetado) ou não há nenhuma solução. Em ambas as situações o método MLR falha em produzir uma

resposta, a menos que as inconsistências sejam eliminadas (colunas linearmente independentes da matriz **X** sejam selecionadas). Alternativamente, podem-se utilizar métodos como a regressão em componentes principais (PCR do inglês *Principal Component Regression*) ou a regressão em quadrados mínimos parciais (PLS, do inglês *Partial Least Squares*) que realizam a projeção do vetor **y** num subespaço gerado pelas componentes obtidas na decomposição de **X** [21-23].

No método PCR, certo número de componentes principais é obtido a partir da decomposição da matriz **X** utilizando PCA e a matriz de escores desta decomposição é utilizada na equação de regressão para representar a matriz **X** de dados original. Desse modo, a variável dependente é projetada num subespaço (com dimensão igual ao número de componentes principais) onde as "novas" variáveis são não correlacionadas, pois formam uma base ortogonal (veja a descrição da PCA na seção 1.3). Portanto, os problemas citados acima são superados e espera-se que as componentes principais descrevam bem a variável dependente. Entretanto, como os escores da decomposição são obtidos utilizando apenas a informação de **X**, a relação dos mesmos com a variável dependente pode não ser a melhor possível, uma vez que na determinação do subespaço de componentes principais as informações contidas nesta última não são consideradas. Assim, uma base ortogonal mais apropriada seria aquela que descreve bem tanto a matriz **X** quanto o vetor **y**. É exatamente essa base que é procurada no método PLS [24,25].

1.7. Quadrados Mínimos Parciais

A regressão por quadrados mínimos parciais (PLS) é atualmente o método de calibração multivariada mais popular na química, mas originalmente foi desenvolvido no campo da econometria por Herman Wold, em meados de 1975. A versão quimiométrica da regressão por PLS foi originalmente proposta por Svante Wold em

1983 como um algoritmo em dois blocos, denominado NIPALS (do inglês *Non Iterative Partial Least Squares*), consistindo de uma sequência de simples modelos parciais ajustados por quadrados mínimos [22,24].

Conforme já comentado, o método PLS define os vetores das base designadas como variáveis latentes de modo a manter um compromisso entre ajustar **X** e predizer **y**. Devido a diferentes visões sobre esse "compromisso" surgiram, desde a proposta de Svante Wold, diferentes algoritmos para a execução do PLS, todos, porém, virtualmente alcançando a mesma solução. No entanto, na proposição original a obtenção das componentes preza por maximizar a covariância entre os escores de **X** e **y,** e assim levar em conta a informação contida em **y** [21-25]. A literatura contém uma grande lista de publicações tratando dos diversos algoritmos utilizados para construir modelos PLS (veja as referências [21] e [22] e os trabalhos nelas citados).

O número ótimo, *R*, de variáveis latentes (VL) do modelo PLS deve ser escolhido antes de usar o modelo para a predição de amostras não pertencentes ao conjunto de calibração inicial. Este parâmetro deve proporcionar ao modelo uma descrição satisfatória das principais características nos dados, ou seja, uma compressão apropriada, resolvendo a questão da colinearidade das variáveis. Vale ressaltar que uma maior compressão, associada a um menor número de VL, faz com que os modelos sejam cada vez menos livres de viés (*bias*). Por outro lado, o aumento do número de VL diminui o viés dos modelos, mas pode levá-los a problemas de sobreajustes. Portanto, estes pontos fazem da escolha do número ótimo de VL uma etapa muito importante e crítica na construção de modelos PLS, devendo existir um equilíbrio entre a não obtenção de modelos subajustados (alta compressão e alto viés) e sobreajustados (baixa compressão e baixo viés).

Um dos procedimentos mais utilizados para a escolha do número de VL é a validação cruzada pela metodologia *leave-N-out*. Nesta abordagem, a partir do conjunto de dados original, um número *N* de amostras é deixado fora do conjunto de calibração,

sendo construídos modelos com diferentes números de variáveis latentes para as amostras restantes. Estes modelos são então usados para predizer a propriedade modelada para as amostras removidas e para as amostras de calibração, e assim alguns valores estatísticos são calculados comparando os valores calculados com os valores verdadeiros (ou observados), como, a raiz quadrada do erro médio da validação cruzada (RMSECV, do inglês *Root Mean Squared Error of Cross Validation*), a raiz quadrada do erro médio da calibração (RMSEC, do inglês *Root Mean Square Error of Calibration*), o coeficiente de correlação da validação cruzada (Q^2), o coeficiente de determinação múltipla da calibração (R^2), entre outros. Estes valores são usados como diagnóstico dos modelos, com o intuito de determinar o número ótimo de variáveis latentes evitando sobreajuste [21,26]. Em grande parte dos trabalhos, o valor de *N* é igual a um, porém, com o objetivo de verificar a robustez do modelo o valor de *N* pode ser aumentado até certa porcentagem do tamanho da matriz original [26].

Uma vez determinado o número de variáveis latentes de um modelo PLS é ainda necessário avaliar a sua capacidade preditiva, que é acessada através da validação externa. Nesta validação o conjunto de dados é dividido em um conjunto de treinamento, que servirá para a construção do modelo, e um conjunto de validação externa (ou de teste), usado para a verificação da capacidade de predição do modelo para amostras que não foram incluídas na obtenção da relação matemática. À exemplo da validação cruzada, parâmetros diagnósticos são calculados a partir da predição para as amostras do conjunto de validação externa, comparando-os com os valores observados para as amostras de testes, como, a raiz quadrada do erro médio da predição (RMSEP, do inglês *Root Mean Squared Error of Prediction*) e o coeficiente de determinação múltipla da predição (R^2_{ext}). As equações para o cálculo dos parâmetros RMSECV, RMSEC, RMSEP, Q^2, R^2 e R^2_{ext} são mostradas na Tabela 1, sendo desejáveis para os três primeiros valores baixos, enquanto para os três últimos valores próximos de um.

1.8. Análise Discriminante por Quadrados Mínimos Parciais

A análise discriminante por quadrados mínimos parciais (PLS-DA, do inglês *Partial Least Squares Discriminant Analysis*) corresponde a um uso específico do método PLS, onde a propriedade modelada (variável dependente) é um vetor **y** especial que contém as informações de classes das diferentes amostras [27]. Através da modelagem deste vetor o método fornece o subespaço (com dimensão igual ao número de variáveis latentes, *R*), no qual as projeções das amostras de diferentes classes têm o máximo de diferenciação. Desse modo, o método PLS-DA é geralmente utilizado para fornecer o conhecimento sobre as variáveis discriminatórias que em algumas aplicações podem ser: biomarcadores que distinguem um grupo controle de um grupo submetido a um tratamento [28,29]; descritores moleculares que distinguem compostos com alta atividade biológica de outros com baixa ou nenhuma atividade; marcadores químicos para a distinção de amostras com diferentes atributos sensoriais [27], entre outras.

No que diz respeito à variável dependente, no método PLS-DA, para especificar uma amostra do conjunto de treinamento pertencente à classe de interesse, o número um (1) é usado, enquanto o número zero (0) indica uma amostra pertencente a uma classe diferente. Alternativamente, quando apenas duas classes estão presentes, pode ser empregado na calibração um vetor com valores +1 para uma classe e -1 para a outra. A classificação de novas amostras é obtida a partir dos valores preditos pelo modelo PLS que não são necessariamente +1 ou -1 (ou 0), mas devem idealmente estar próximo dos valores utilizados para identificar as classes [30-33].

A tradução dos valores preditos para cada classe (ou seja, o procedimento de classificação propriamente dito) é um ponto crítico na análise por PLS-DA e pode ser feita, por exemplo, aplicando um valor limite (*threshold*) acima do qual a amostra será assinalada à classe de interesse (valor +1) e abaixo para as demais classes (valores 0 ou -1). Quando os valores +1 e 0 são usados, uma abordagem simples seria definir

arbitrariamente o *threshold* como sendo 0,5 ou quando os valores +1 e -1 são usados, definir o valor 0 como *threshold*. Na literatura diversas abordagens mais avançadas foram propostas envolvendo pressuposições sobre a distribuição estatística dos valores preditos; o estabelecimento de faixas de incerteza para a classificação; a consideração da probabilidade de cada decisão de classificação, entre outras [34,35].

Os valores de falsos-positivos (amostras que pertencem à classe -1 ou 0 que são classificadas na classe de interesse +1) e falsos-negativos (amostras que pertencem à classe de interesse e que são classificadas na classe -1 ou 0) podem ser organizados juntos dos valores de verdadeiros-positivos (amostras da classe de interesse classificados corretamente) e verdadeiros-negativos (amostras da classe -1 ou 0 classificados corretamente) numa matriz de confusão. Um exemplo de uma matriz de confusão para um problema de classificação do tipo +1/-1 é mostrado na Figura 3. A partir dos valores na matriz citada, as figuras de mérito seletividade e sensibilidade podem ser calculadas. A seletividade é a razão entre o número de verdadeiros-negativos e a soma de verdadeiros-negativos com falsos-positivos e a sensibilidade é a razão entre o número de verdadeiros-positivos e a soma de verdadeiros-positivos com falsos-negativos. A sensibilidade é uma medida da habilidade do modelo para classificar corretamente as amostras da classe de interesse, enquanto a seletividade mede a habilidade do modelo para predizer as amostras das outras classes.

Tabela 1: Parâmetros estatísticos utilizados como diagnóstico para a qualidade de modelos PLS.

Parâmetro	Definição[a]
Raiz quadrada do erro médio da validação cruzada	$\text{RMSECV} = \sqrt{\sum_{i=1}^{I} \frac{(y_i - \hat{y}_{vci})^2}{I}}$
Raiz quadrada do erro médio da calibração	$\text{RMSEC} = \sqrt{\sum_{i=1}^{I} \frac{(y_i - \hat{y}_{ci})^2}{I - R - 1}}$
Raiz quadrada do erro médio da predição	$\text{RMSEP} = \sqrt{\sum_{i=1}^{M} \frac{\left(y_i - \hat{y}_{pi}\right)^2}{M}}$
Coeficiente de correlação da validação cruzada	$Q^2 = 1 - \sum_{i=1}^{I} \frac{(y_i - \hat{y}_{vci})^2}{(y_i - \bar{y}_i)^2}$
Coeficiente de determinação múltipla da calibração	$R^2 = 1 - \sum_{i=1}^{I} \frac{(y_i - \hat{y}_{ci})^2}{(y_i - \bar{y}_i)^2}$
Coeficiente de determinação múltipla da predição	$R^2_{ext} = 1 - \sum_{i=1}^{M} \frac{\left(y_i - \hat{y}_{pi}\right)^2}{(y_i - \bar{y}_i)^2}$

[a] y_i – valor observado da propriedade modelada para a *i*-ésima amostra e $\bar{y}_i$ é média aritmética destes valores; $\hat{y}_{vci}$ – valor calculado na validação cruzada da propriedade modelada para a *i*-ésima amostra; $\hat{y}_{ci}$ – valor calculado na calibração da propriedade modelada para a *i*-ésima amostra; $\hat{y}_{pi}$ – valor calculado da propriedade modelada para a *i*-ésima amostra no conjunto de validação externa; *I* é o número de amostras no conjunto de calibração; *R* é número de variáveis latentes do modelo PLS e *M* é o número de amostras no conjunto de validação externa.

Como alternativa à escolha de um único *threshold*, um "espectro" do desempenho do modelo PLS-DA para a faixa inteira de *thresholds* pode ser avaliado. Esse "espectro" é conhecido como curva da característica do operador receptor (ROC, do inglês *Receiver Operator Characteristic*) e é obtido pelo gráfico de 1-seletividade *versus* a sensibilidade

para cada valor de *threshold* dentro do intervalo entre os valores utilizados para definir as classes. Por exemplo, para a definição da ROC num problema de classificação como aquele da Figura 3, a faixa de *thresholds* é limitada pelos valores de +1 e -1, sendo utilizado algum incremento para definir os valores possíveis. O valor de 1-seletividade corresponde à taxa de falsos-positivos (razão entre o número de falsos-positivos e a soma de verdadeiros-negativos com falsos-positivos) e a sensibilidade corresponde à taxa de verdadeiros- positivos.

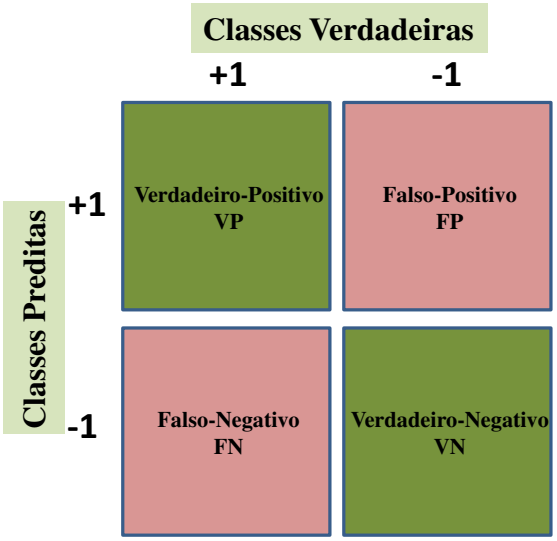

Figura 3: Matriz de confusão para um problema de classificação +1/-1.

Considerando um problema de classificação entre duas classes (um delas sendo a classe de interesse) com o mesmo número de amostras, um modelo PLS-DA inútil seria aquele onde as taxas de verdadeiros-positivos e falsos-positivos são sempre iguais entre si, independente do *threshold*, ou seja, este modelo prediz todas as amostras na classe de interesse (veja os quadros superiores na Figura 3). Assim, a probabilidade de uma amostra do conjunto ser realmente da classe de interesse é de 50%. Esta situação é mostrada na Figura 4 através da curva ROC representada por uma linha vermelha que passa pela diagonal do quadrado de lado igual a um. A área sob a ROC em questão é exatamente igual à probabilidade citada (0,50 ou 50%) e constitui um parâmetro

diagnóstico do poder discriminatório de um modelo de classificação. Deste modo, a área sob a característica do operador receptor (AUROC, do inglês *Area Under the Receiver Operator Characteristic*) pode ser determinada para diferentes modelos, no intuito de avaliar suas qualidades.

Na Figura 4 outras duas curvas ROC são representadas, sendo aquela de cor azul representativa de um melhor modelo do que o modelo daquela de cor verde. Observa-se para a curva ROC azul maiores taxas de verdadeiros-positivos e assim AUROC mais próxima de um. Uma AUROC igual a um representa um modelo com classificação perfeita, enquanto, conforme já citado, uma AUROC igual a 0,50, um modelo de classificação inútil [30].

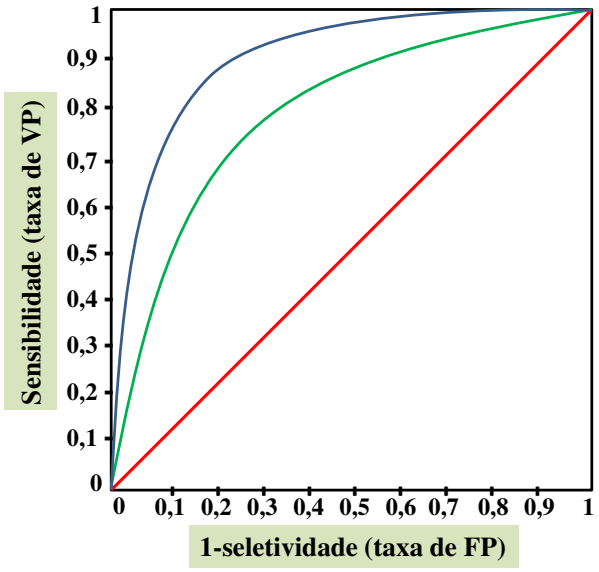

Figura 4: Representação de três curvas ROC. A curva ROC azul representa o melhor modelo (AUROC mais próximo de um) e a curva ROC vermelha exemplifica um modelo de classificação inútil. A curva ROC verde mostra um modelo intermediário.

Outros valores estatísticos podem também ser usados como diagnóstico para os modelos PLS-DA, como por exemplo, o coeficiente de correlação na validação cruzada Q^2, o número de erros na classificação (NMC, do inglês *Number of Misclassifications*), entre outros. O NMC é considerado o diagnóstico estatístico mais intuitivo, uma vez que

simplesmente indica o número de amostras que são classificadas erradamente pelo modelo. Em outras palavras, o NMC é a soma dos falsos-positivos com os falsos-negativos [30].

Por fim, vale ressaltar que o método PLS-DA deve ser utilizado com bastante cuidado, tendo em vista a relativa facilidade de serem obtidos modelos de classificação ao acaso, ou seja, modelos que apesar de mostrarem boa capacidade discriminatória foram obtidos puramente por coincidência. Isto acontece, pois em geral na química os conjuntos de dados possuem muitas variáveis para poucas amostras, quando seria necessário um número consideravelmente maior de amostras para descrever precisamente o problema de classificação. Além disso, as propriedades dos algoritmos para execução do PLS fazem dele um método muito eficaz para encontrar o subespaço que fornece a discriminação, mas que nem sempre resiste a uma exigente validação [32].

Para verificar se um modelo PLS-DA foi obtido por chance, uma boa opção é realizar um teste de hipótese utilizando permutações. Neste teste, é admitida uma hipótese nula, H_0, de que não existe diferença entre as classes contra uma hipótese alternativa, H_1, de que existe diferença entre as classes e que esta é traduzida pelo modelo PLS-DA que está sendo testado. Assim, tendo em vista H_0, o "rótulo" de classes presente no vetor **y** pode ser atribuído de forma aleatória entre as amostras. Deste modo, o vetor de classes original tem seus valores permutados, sendo construído um novo modelo PLS-DA, mantendo o mesmo número de VL do modelo original. Como muitas permutações são possíveis, este processo é repetido muitas vezes e a cada uma destas permutações, um modelo diferente é obtido. Através do cálculo de um parâmetro estatístico, como por exemplo, o NMC, para cada um destes modelos, uma distribuição para hipótese nula (H_0) é obtida calculando as frequências de ocorrência dos parâmetros.

Se o parâmetro estatístico do modelo PLS-DA original não é significativamente diferente daqueles da distribuição de H_0, ou seja, ele faz parte da mesma distribuição, H_0 é aceita para o modelo (rejeitando-se H_1) e assim o mesmo é considerado como sendo

obtido por chance. No caso contrário, sendo o parâmetro estatístico do modelo original significativamente diferente, H_0 é rejeitada (aceitando-se H_1) e a inferência estatística para diferença entre as classes não é considerada como obtida ao acaso.

A significância da diferença entre o parâmetro estatístico do modelo original e a distribuição de H_0 é obtida calculando-se o valor de *p* estatístico e comparando-o a um nível de significância (α) que em geral é de 5% ou 0,05. Se em relação a um dado parâmetro estatístico um menor valor está associado a um melhor desempenho do modelo original (por exemplo, o NMC), espera-se, para rejeição de H_0, que os modelos das permutações possuam valores deste parâmetro significativamente maiores (desempenho estatístico inferior). Em outras palavras, há uma comparação entre o modelo original e a cauda à esquerda da distribuição de H_0. Para este caso, o valor de *p* estatístico é calculado como sendo o número de modelos com qualidade igual ou superior ao modelo testado (valor de NMC menor) mais 1, dividido pelo número de permutações realizadas. Neste contexto, a Figura 5A mostra um exemplo, onde a hipótese nula não pode ser rejeitada para o modelo, a um nível de significância $\alpha = 0,05$. No exemplo da Figura 5A, o valor do parâmetro do modelo original é destacado por um ponto vermelho e existem 1559 modelos da distribuição de H_0 com valores menores, sendo 10.000 permutações realizadas, assim, o valor de *p* para o teste é 0,156 e como este é maior do que o nível de significância ($\alpha = 0,05$), a hipótese nula não pode ser rejeitada e o modelo é obtido ao acaso. Por outro lado, na Figura 5B o valor de *p* do teste é 0,021 e assim como este é menor do que a significância pode ser dito que o modelo não é obtido ao acaso e tem classificação significativa no nível testado ($\alpha = 0,05$) [30-32].

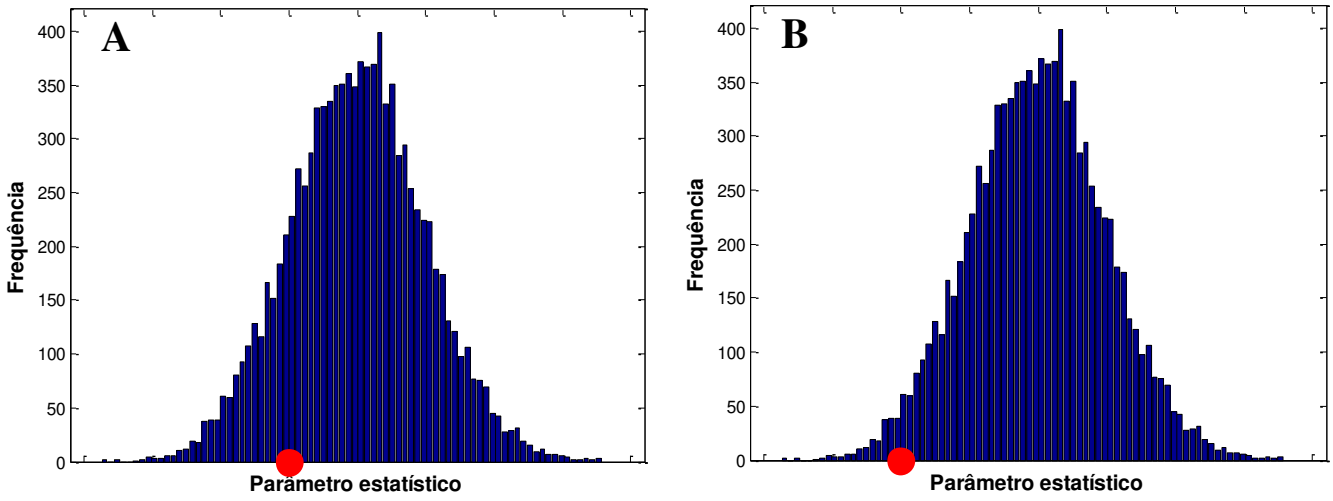

Figura 5: Exemplos de testes de aleatorização para um modelo com um parâmetro estatístico (mostrado através de um ponto vermelho) comparado à cauda à esquerda de uma distribuição de hipótese nula (H_0) incluindo 10.000 modelos aleatorizados e em um nível de significância de $\alpha = 0,05$. Neste exemplo, quanto menor o parâmetro estatístico melhor é o modelo. A: A hipótese nula não pode ser rejeitada (classificação não significativa = modelo ao acaso), pois o valor de *p* é igual a 0,156; B: A hipótese nula pode ser rejeitada (classificação significativa = modelo não é ao acaso), pois o valor de *p* é igual a 0,021.

Para o teste utilizando parâmetros como, a AUROC ou Q^2, onde maiores valores são desejáveis, a avaliação é feita pela comparação com a cauda à direita da distribuição e assim o valor de *p* é calculado observando os modelos aleatorizados com maiores valores destes parâmetros do que aquele para o modelo original. Vale ressaltar que no teste citado o número de aleatorizações é importante e em geral deve ser o maior possível, para que muitas permutações sejam realizadas e assim haja uma melhor amostragem das caudas da distribuição [30-32].

Capítulo 2: *Bucketing* otimizado para espectros de ressonância magnética nuclear: três estudos de caso

2.1. Introdução

A espectroscopia de ressonância magnética nuclear (RMN) é uma técnica poderosa, versátil, não destrutiva, não invasiva e bastante reprodutível. Além de ser adequada para elucidação estrutural, a RMN pode ser usada para analisar amostras complexas sem a necessidade de separação física. Estas vantagens podem ser exploradas por ferramentas quimiométricas apropriadas fornecendo vários tipos de informações úteis no reconhecimento de padrões, na detecção de adulterações, em perfis metabólicos, entre outras aplicações [36-40].

A exploração da informação química "codificada" nos espectros de RMN é prejudicada pelos desalinhamentos frequentes, especialmente nos espectros de RMN de hidrogênio (RMN ^{1}H). Estes surgem devido a vários fatores, tais como, instabilidade instrumental, pH, força iônica, temperatura, entre outros, e podem conduzir as análises estatísticas a interpretações incorretas dos resultados, pois a pressuposição da bi- ou multilinearidade das abordagens quimiométricas, sobre a natureza dos dados, não são consideradas apropriadamente [41-44].

Na literatura, diferentes métodos foram propostos para corrigir os desalinhamentos. Uma abordagem popular e de baixo custo computacional é chamada de *bucketing* [41]. Alternativamente, métodos mais elaborados têm sido empregados, tais como, a deformação otimizada por correlação (COW, do inglês *Correlation Optimized Warping*) [45-49], a deformação temporal dinâmica (DTW, do inglês *Dynamic Time Warping*) [45,47], a deformação por correlação (*coshift*, do inglês *correlation shifting*) [50], a deformação por correlação em intervalos (*icoshift*, do inglês

interval correlation shifting) [42,50]. O método *bucketing* é teoricamente mais simples do que as outras abordagens citadas.

Na verdade, o método *bucketing* realiza uma redução nos dados através do agrupamento das respostas espectrais, funcionando como um método de alinhamento dos dados. No método convencional, os espectros são divididos em janelas igualmente espaçadas, denominadas *buckets*, e cuja largura comumente varia entre 0,01 a 0,05 ppm. As intensidades dentro de cada *bucket* são somadas, tal que a área sob cada região espectral é usada em vez de intensidades individuais. Portanto, um novo conjunto de variáveis (onde cada uma delas é o resultado da soma de intensidades dentro de cada *bucket*) é gerado e como a largura dos *buckets* é definida com o intuito de cobrir a variabilidade de deslocamentos químicos ao redor dos picos, o desalinhamento tende a ser resolvido [41,51,52].

A Equação 13 resume o procedimento de *bucketing* aplicado a uma matriz de dados $\mathbf{X}$ ($I \times J$) com I amostras e J variáveis, onde z_{ik} corresponde ao sinal da i-ésima amostra resultante do k-ésimo *bucket*, obtido após o procedimento.

$$z_{ik} = \sum_{j=N*(k-1)+1}^{N*k} x_{ij} \qquad k = 1, 2, \ldots, K \qquad \text{Equação 13}$$

Para cada amostra i, x_{ij} é a intensidade do sinal original no ponto j. O parâmetro N é o número de pontos de dados em cada *bucket* e pode ser calculado pela razão entre a largura do *bucket* e o intervalo de amostragem no espectro, que corresponde ao incremento entre dois pontos subsequentes no espectro. Por exemplo, se a largura do *bucket* é 0,05 ppm e o intervalo de amostragem é 0,0005 ppm, o parâmetro N será igual a 0,05/0,0005 = 100 pontos. Este intervalo varia entre diferentes experimentos, pois depende de como a aquisição das curvas de decaimento induzido livre (FID do inglês *Free Induced Decay*) é realizada, ou seja, depende dos parâmetros definidos no experimento de RMN, tais como, tempo de aquisição, pontos de dados totais

digitalizados e largura espectral. O parâmetro *K* é o número final de *buckets* e é igual à parte inteira da razão *J*/*N*. Portanto, de acordo com a Equação 13, um novo eixo *k* no domínio das variáveis é gerado, onde as intensidades z_{ik} são organizadas numa matriz **Z** (*I* × *K*) com *I* amostras e *K buckets*. Vale ressaltar que quando o produto *NK* não é igual a *J*, o último *bucket* na matriz **X** é menor do que os demais, de modo que as *J* variáveis sejam incluídas no procedimento. Ainda, se for o caso, para contornar efeitos devido às diferenças entre os volumes de amostra utilizados no experimento de RMN, cada espectro na matriz **Z** pode ser normalizado, sendo cada variável dividida pela soma de todas as outras no espectro, tornando assim a soma total igual a um e os espectros com área total unitária.

Uma desvantagem deste método é que um mesmo sinal de ressonância pode aparecer em dois ou mais *buckets*, dividindo a informação química em questão. Isto ocorre porque o *bucketing* convencional usa fronteiras rígidas. Apesar disso, vários trabalhos na literatura [36,37,53,54] usam efetivamente esta metodologia.

A Figura 6 mostra um conjunto simulado de espectros de RMN desalinhados. As Figuras 6B e 6C ilustram o procedimento de *bucketing* convencional. Como pode ser visto na Figura 6B, que apresenta a média dos espectros simulados com as fronteiras dos *buckets* assinaladas por linhas verticais, o procedimento com largura de *bucket* de 0,01 ppm é incapaz de isolar apropriadamente os sinais. Como resultado, na Figura 6C, onde a soma das intensidades para cada *bucket* é apresentada através de barras coloridas (as alturas das barras são relacionadas aos valores das integrais que foram normalizadas para a soma total igual a um), cinco variáveis importantes são observadas, contendo a informação principal do conjunto de dados, que na verdade tem três sinais. Isto pode, consequentemente, dificultar seriamente as interpretações, por exemplo, quando uma análise de componentes principais é utilizada.

A desvantagem citada acima pode ser superada através do uso de fronteiras de *buckets* ajustadas aos mínimos locais, com o objetivo de fornecer *buckets* otimizados de

diferentes tamanhos. De fato, um tipo de solução similar já foi proposto na literatura, como por exemplo, a metodologia para *bucketing* implementada no *software* comercial ACD/Labs® (Toronto, Canadá) denominada *intelligent bucketing* [51,52]. Neste método, o programa escolhe as divisões baseando-se nos mínimos locais, assim procurando pelo melhor modo de dividir o espectro e evitar o problema do método convencional. Entretanto, o programa não possui código aberto e o método para encontrar os mínimos não foi reportado.

Em outro trabalho, Davis *et al.* [55] propuseram uma metodologia denominada *adaptive binning*, onde a transformada *wavelet* sem o procedimento de subamostragem (divisão pela metade do conjunto de dados a cada nível de decomposição) é usada para eliminar o ruído e encontrar todos os mínimos em um espectro de referência (espectro composto pelo valor máximo de cada variável entre todas as amostras) realizando a seguir, a soma das intensidades entre estes mínimos, para cada espectro da matriz original. Entretanto, na decomposição a predefinição dos níveis e do tipo de função de base *wavelet* é necessária. Assim, existe uma dependência no número de níveis na decomposição, além do *threshold* para a eliminação do ruído. No capítulo 3, a seguir, será comentado brevemente sobre a teoria da transformada *wavelet* com aplicação da mesma para eliminação de ruído (*denoising*).

Alternativas para o *bucketing* tradicional foram propostas na literatura recentemente, denominadas *gaussian binning* [56] e *dynamic adaptive binning* [57], mas da mesma forma que o método proposto por Davis *et al.* [55], estas metodologias requerem um alto nível de *expertise* do usuário, sendo mais complexas que o algoritmo que será aqui apresentado.

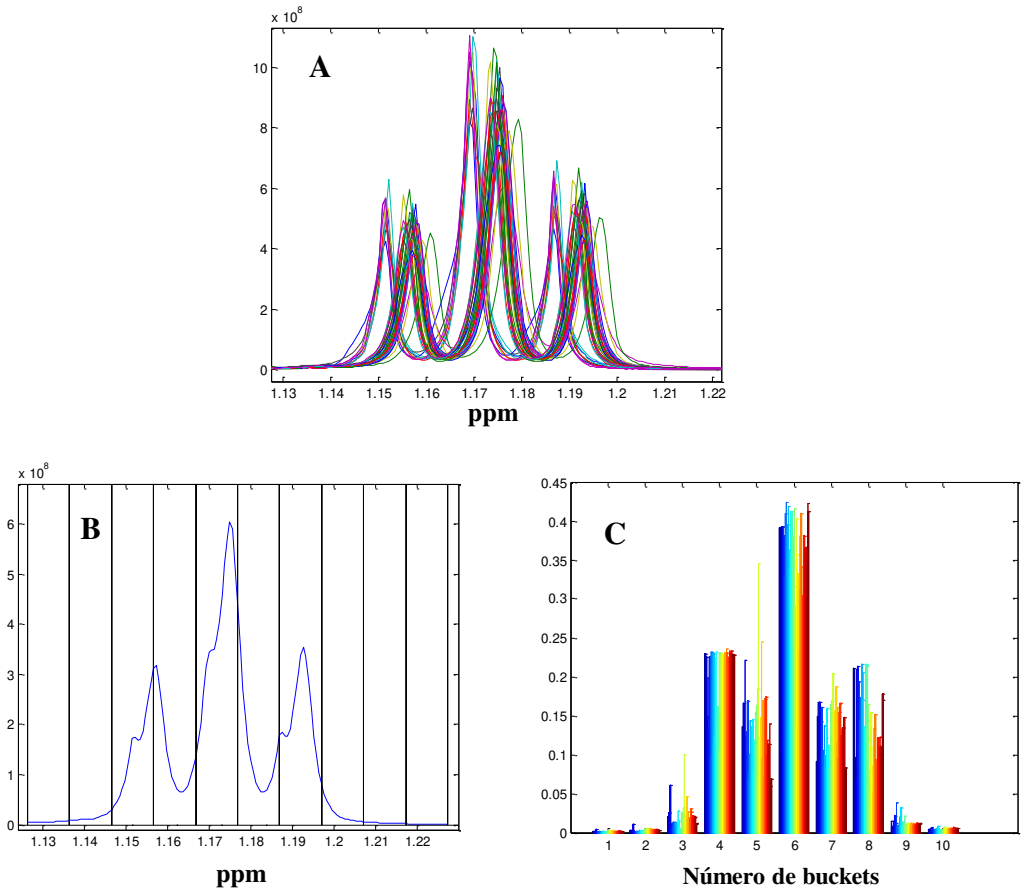

Figura 6: Esquema para o procedimento de *bucketing* convencional. A: Espectros de RMN simulados com desalinhamentos. B: Espectro simulado médio e as fronteiras dos *buckets* (linhas verticais) delimitadas pelo *bucketing* convencional, com tamanhos de *buckets* de 0,01 ppm. C: Valores normalizados nos *buckets* de cada amostra apresentados através de um gráfico de barras obtido pelo *bucketing* convencional (as cores representam as diferentes amostras).

2.2. Objetivo

Neste capítulo, o objetivo é apresentar um método para corrigir desalinhamentos, que otimiza os tamanhos dos *buckets* através da definição das suas fronteiras em mínimos locais, determinados no espectro de RMN médio ou um espectro de referência. A metodologia é designada "*optimized bucketing algorithm*" (OBA) ou algoritmo de *bucketing* otimizado, sendo simples, acessível a qualquer usuário como uma rotina escrita em código Matlab (arquivo .m) e disponível para *download* gratuito em

http://lqta.iqm.unicamp.br. Por fim, o método OBA é aplicado a conjuntos de dados e comparado a outros métodos da literatura.

2.3. Metodologia

2.3.1. O "*optimized bucketing algorithm*"

O método "*optimized bucketing algorithm*" (OBA) que está sendo proposto é uma modificação do procedimento convencional. Com o intuito de definir *buckets* com tamanho variável, mas comuns a todas as amostras, um espectro de referência que pode ser o espectro médio $\bar{\mathbf{x}}^{\mathrm{T}}$ é usado, onde cada elemento $\bar{x}_j$ é a média da *j*-ésima coluna de **X** (*I* × *J*). Primeiramente, dois parâmetros são definidos: 1) o tamanho inicial dos *buckets* em ppm, que é convertida em número de pontos *N* pelo algoritmo, usando o intervalo de amostragem calculado a partir do eixo de deslocamentos químicos em ppm e 2) a flexibilidade, que é dada como uma porcentagem de *N* e define o quanto as fronteiras podem se mover na busca pelos mínimos locais no espectro médio. Ambos os parâmetros são definidos *a priori* e necessários para a execução do algoritmo. O valor de flexibilidade é convertido, no algoritmo, no parâmetro *s*, expresso como flexibilidade*0,01**N* (o valor 0,01 surge para mudar a flexibilidade de porcentagem para fração). Portanto, usando como entrada a largura inicial de *buckets* de 0,04 ppm, por exemplo, em um conjunto de dados com intervalo de amostragem de 0,0004 ppm (*N* = 0,04/0,0004 = 100 pontos) e flexibilidade de 50% (*s* = 50*0,01*100 = 50 pontos), o tamanho dos *buckets* pode variar de 0,02 até 0,06 ppm (100 ± 50 pontos), dependendo de onde o mínimo local é encontrado. As saídas do algoritmo são: a matriz resultante com dimensões (*I* × *K*), onde *K* é o número total de *buckets*, as fronteiras otimizadas e o tamanho resultante de cada *bucket*, ambos em ppm.

Do ponto de vista matemático, o método OBA pode ser entendido como segue: uma vez que o tamanho inicial dos *buckets* e a flexibilidade são conhecidos, o vetor $\mathbf{v}^{\mathrm{T}}$ (Equação 14), cujos elementos definem as fronteiras dos *buckets*, é gerado. O primeiro *bucket* inicia na variável $j = 1$ e o último termina na variável $j = J$ do espectro médio, $\bar{\mathbf{x}}$, e estes são o primeiro e o último elemento do vetor $\mathbf{v}^{\mathrm{T}}$. Os outros elementos deste vetor são, na verdade, os índices q dos q-ésimos elementos de $\bar{x}_q$ que correspondem aos mínimos locais na região delimitada por $\bar{x}_{N*t-s}$ e $\bar{x}_{N*t+s}$ (esta região é definida com o auxílio do símbolo "**:**" na Equação 15, conforme linguagem do Matlab), onde $t = 1, 2, \ldots, T$, com T sendo igual à parte inteira de $(J/N) - 1$, conforme definido na Equação 15 (nesta equação o símbolo "min" significa o menor valor dentre todos os elementos do intervalo delimitado pelos elementos indicados).

$$\mathbf{v}^{\mathrm{T}} = [1, \cdots, q, \cdots, J] \qquad \text{Equação 14}$$

$$\bar{x}_q = \min(\bar{x}_{N*t-s} : \bar{x}_{N*t+s}) \qquad \text{Equação 15}$$

Os elementos de $\mathbf{v}^{\mathrm{T}}$ que indicam as posições dos mínimos substituem os limites dos *buckets* que são mostrados na Equação 13, realizando a otimização do procedimento de correção de desalinhamento, para cada amostra i, como mostrado na Equação 16, onde v_k é o k-ésimo elemento do vetor $\mathbf{v}$. A nova matriz $\mathbf{Z}$ ($I \times K$) é obtida, onde as novas intensidades z_{ik} são organizadas e o novo eixo k no domínio das variáveis é gerado.

As Figuras 7B e 7C mostram o resultado obtido quando o procedimento OBA é aplicado ao mesmo conjunto de dados simulados comentados anteriormente (espectros desalinhados). O número de *buckets* K é o mesmo de antes (Figura 6), mas é visível a partir da Figura 7B que o novo algoritmo é capaz de definir as fronteiras nos mínimos locais (linhas verticais). Em ambos os lados do pico central, os *buckets* se tornam mais estreitos, pois na busca pelos mínimos locais, as fronteiras, inicialmente definidas em posições não otimizadas, tendem a se mover para mais próximo das regiões entre os

picos. Como resultado, na Figura 7C, onde a soma das intensidades de cada *bucket*, em cada amostra, são mostrados através de barras coloridas, somente três importantes *buckets* são observados, conforme esperado, uma vez que os dados simulados têm somente três picos. Deste modo, o desempenho superior da metodologia proposta sobre o *bucketing* convencional é visível.

$$z_{ik} = \sum_{j=v_k}^{v_{k+1}} x_{ij} \qquad \text{para } k = 1$$

$$z_{ik} = \sum_{j=v_k+1}^{v_{k+1}} x_{ij} \qquad \text{para } k = 2, \ldots, K = length(\mathbf{v}) - 1$$

Equação 16

Uma questão importante a ser considerada no método OBA é a escolha da melhor combinação entre o tamanho inicial dos *buckets* e a flexibilidade para cada conjunto de dados. A inspeção visual da extensão dos desalinhamentos na linha de base pode ser de grande ajuda na definição destes parâmetros. Também, algum critério, como por exemplo, a variância explicada nas primeiras componentes principais em uma PCA da matriz resultante $\mathbf{Z}$ ($I \times K$) pode dar uma estimativa razoável da combinação entre os valores de entrada.

Finalmente, não é aconselhável o uso de *buckets* iniciais muito grandes, pois esta abordagem promove uma diminuição na resolução espectral, devendo existir assim um compromisso entre o ganho com a correção dos desalinhamentos e a redução no número de variáveis. Para as aplicações apresentadas neste trabalho, os parâmetros de entrada foram determinados pela inspeção visual dos desalinhamentos na linha de base e observação dos gráficos das matrizes obtidas, com o objetivo de escolher as combinações com menor diminuição na resolução espectral.

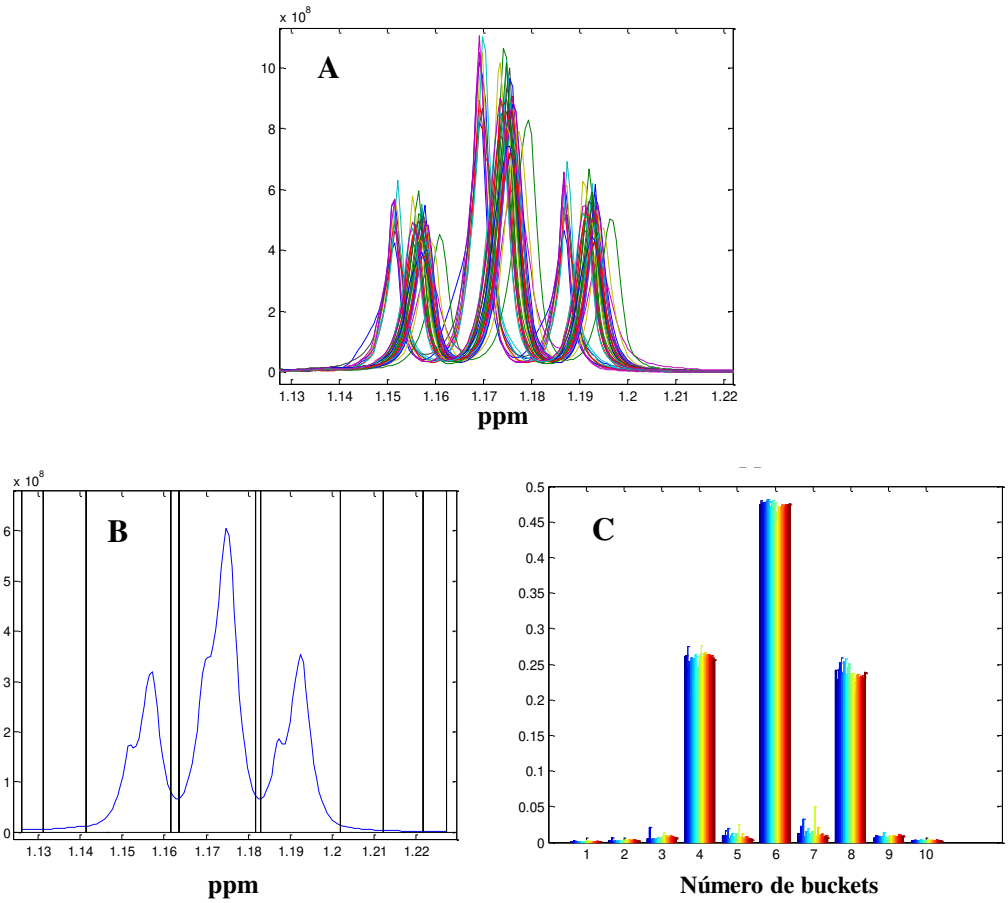

Figura 7: Esquema para o procedimento de *bucketing* otimizado (OBA). A: Espectros de RMN simulados com desalinhamentos. B: Espectro simulado médio e as fronteiras dos *buckets* (linhas verticais) delimitadas pelo OBA, com tamanhos iniciais de *buckets* de 0,01 ppm e flexibilidade de 50%. C: Valores normalizados em cada *bucket* de cada amostra apresentados através de um gráfico de barras obtido pelo OBA (as cores representam as diferentes amostras).

2.4. Aplicações

Para testar a aplicabilidade do método OBA e compará-lo com outros métodos da literatura, três conjuntos de dados de RMN foram selecionados, sendo descritos a seguir. Dois destes conjuntos de dados foram retirados da literatura (conjunto de dados de vinhos e de tumores cerebrais) e o outro foi adquirido no nosso laboratório (conjunto de dados de misturas de biodiesel-diesel).

2.4.1. Parte experimental

2.4.1.1. Conjunto de dados de vinhos

Os espectros de RMN ^{1}H de vinhos foram estudados por Larsen *et al.* [49], onde a metodologia para aquisição dos dados foi descrita. A matriz de dados **X** (40 × 8712), adquirida em http://www.models.kvl.dk/Wine_NMR, é composta de quarenta amostras de vinhos de mesa (distribuídos entre branco, tinto e *rosé*) cobrindo a região de 0,50 a 6,00 ppm, e o conteúdo de ácido lático (um ácido orgânico importante para o perfil de sabor do vinho [49]) de cada amostra. Nesta região, são observados vários picos atribuídos ao etanol, ácidos orgânicos, carboidratos e em menor quantidade, polifenóis, outros compostos aromáticos e corantes. Todos esses picos apresentam desalinhamentos que ocorrem principalmente devido à diferença no pH das amostras, que não foram ajustados antes das análises. O método OBA foi aplicado aos espectros de RMN usando 0,05 ppm como largura inicial dos *buckets* e flexibilidade de 50%. Para este conjunto de dados o intervalo de amostragem foi de cerca de 0,00063 ppm levando a N = 0,05/0,00063 = 79 pontos e s = 50*0,01*79 = 39 pontos (estes parâmetros foram arredondados para o inteiro mais próximo). Como resultado, a matriz de *buckets* **Z** com dimensões (40 × 109) foi obtida. *Buckets* com 0,05 ppm foram usados para o procedimento de *bucketing* convencional e neste caso a matriz **Z** foi obtida com dimensões (40 × 110). Uma vez que o conteúdo de ácido lático estava disponível, a matriz de dados original e as matrizes pré-tratadas (pelo *bucketing* convencional e OBA), usando o espectro inteiro e uma região selecionada, foram usados para a construção de modelos PLS e MLR sobre as matrizes centradas na média. Todos os modelos tiveram seus desempenhos preditivos avaliados pela validação cruzada usando o procedimento de *leave-one-out* onde o coeficiente de determinação (R^2, veja equação

na Tabela 1) foi calculado. Para a construção dos modelos rotinas escritas em código Matlab (The MathWorks, Natick, MA, USA) foram usadas.

2.4.1.2. Conjunto de dados de misturas de biodiesel-diesel

Cem amostras de misturas de biodiesel-diesel coletadas em postos de gasolina no estado de São Paulo foram fornecidas pela Central Analítica do Instituto de Química da Universidade Estadual de Campinas. As amostras foram classificadas como metropolitanas (vindas da região metropolitana de Campinas) e não metropolitanas (outras cidades do interior do Estado), de acordo com as suas regiões de comercialização. Os espectros de RMN ^{1}H foram adquiridos em um espectrômetro Bruker Avance DRX400 operando a 400,13 MHz de frequência a temperatura ambiente, usando 550 μL de amostra pura em uma sonda de 5 mm Bruker BBO sem rotação usando uma sequência de pulso de 90° padrão para ^{1}H. A homogeneidade do campo foi obtida por inspeção do formato de uma linha de um espectro de um padrão (0,3% de clorofórmio em acetona-d6). Esta condição de campo foi usada para todas as amostras durante todas as análises. Todos os espectros foram adquiridos com 32768 pontos (32k) no domínio do tempo, 20 ppm de largura de janela espectral e 16 varreduras (*scans*). Os FIDs foram processados com o programa TOPSPIN 2.1 com 65536 pontos (64k), multiplicados por uma função de janela exponencial com largura de linha constante de 0,3 Hz e transformada de Fourier normal. Os espectros finais tiveram suas fases ajustadas, um a um, por inspeção direta. A correção da linha de base foi feita usando uma função linear automática. Todos os espectros foram referenciados usando uma posição de campo digital obtida usando tetrametilsilano (TMS) em acetona a 0 ppm. Os espectros de RMN ^{1}H foram organizados em uma matriz de dados **X** com dimensões (100 × 15850) relativa à região de 0,02 a 10,00 ppm, que foi reduzida em *buckets* pelo modo convencional, usando largura de *buckets* de 0,05 ppm e OBA com flexibilidade de

50% (intervalo de amostragem = 0,00063 ppm, N = 79 pontos, s = 39 pontos), seguida por normalização para área unitária. Os procedimentos de *bucketing* forneceram matrizes **Z** com dimensões (100 × 200) e (100 × 191) para os métodos convencional e otimizado, respectivamente. Os dados originais e as matrizes **Z** foram centrados na média e submetidos à análise exploratória por PCA usando o programa Pirouette 3.11 (Infometrix, Seattle, WA, USA).

2.4.1.3. Conjunto de dados de tumores cerebrais

Os espectros de RMN ^{1}H dos extratos de tumores cerebrais foram estudados por Faria *et al.* [33] onde a metodologia para a aquisição dos dados foi descrita. A partir dos espectros de RMN ^{1}H reportados, 16 e 13 espectros correspondendo aos tumores não neurogliais (NN) e neurogliais em alto grau (Hg), respectivamente, foram selecionados para o presente estudo. Os espectros correspondendo à região entre 1,22 e 4,25 ppm foram organizados em uma matriz de dados **X** com dimensões (29 × 4964). O método OBA foi aplicado aos espectros de RMN usando 0,002 ppm como largura inicial de *buckets* e flexibilidade de 50%. Para este conjunto de dados o intervalo de amostragem foi cerca de 0,00060 ppm levando a N = 0,002/0,00060 = 3 pontos e s = 50*0,01*3 = 2 pontos (estes parâmetros foram arredondados para o inteiro mais próximo). Como resultado, a matriz **Z** com dimensões (29 × 1416) foi obtida. A análise discriminante por quadrados mínimos parciais (PLS-DA) com a abordagem de validação cruzada *leave-one-out* foi usada para construir modelos de classificação (entre os dois tipos de tumores, NN e Hg) a partir do conjunto de dados original e do conjunto de dados após o pré-tratamento com OBA, usando um vetor **y** de classes onde o valor +1 foi definido para os tumores NN e o valor -1 foi definido para os tumores Hg. As análises foram realizadas sobre as matrizes centradas na média. A avaliação se os modelos PLS-DA foram obtidos por chance foi realizada comparando o número de erros na classificação

(NMC), definido como a soma de falsos-positivos (FP) e falsos-negativos (FN), aos NMCs (diagnósticos estatísticos) obtidos de 10.000 testes de permutação, computados usando o vetor **y** de classes permutado (aleatorização de **y** ou em inglês **y**-*randomization*). Os FP e FN são obtidos relacionando as classes preditas a um *threshold* discriminativo definido usando as estimativas das distribuições para os valores preditos em cada classe. O *threshold* foi selecionado no ponto onde as duas distribuições estimadas são iguais, sendo estas distribuições assumidas como aproximadamente normais (distribuições gaussianas com média e desvio padrão das predições para cada classe).

Para os modelos das permutações, uma hipótese nula H_0 assume que não há diferença entre os grupos de tumores. Assim, a significância da classificação dos modelos, ou seja, se os mesmos são ou não obtidos ao acaso, é acessada comparando seus valores de NMC com aqueles referentes aos modelos das distribuições de hipótese nula H_0 [30,32]. A partir destas comparações, cada valor de *p* estatístico (um mais o número de elementos na distribuição nula que são menores ou iguais ao NMC para o modelo não permutado dividido pelo número de testes de permutação, neste caso, 10.000) foi calculado [30] e associado a um nível de significância de $\alpha = 0,05$. Detalhes adicionais sobre esta avaliação de desempenho podem ser encontrados em [30,32]. Todas as análises foram realizadas usando rotinas escritas em código Matlab (The MathWorks, Natick, MA, USA).

2.4.2. Resultados e Discussão

2.4.2.1. Conjunto de dados de vinhos

Neste conjunto de dados uma ampla faixa de deslocamentos de picos é observada, fortemente dependentes do pH da amostra. Isto pode ser visto na Figura 8 nos espectros

de RMN ^{1}H para todas as amostras, nas regiões ampliadas relacionadas aos sinais do etanol (quarteto do grupo metileno em 3,6 ppm) e ácido lático (dubleto do grupo metila terminal em 1,4 ppm). O alinhamento destes espectros usando as abordagens COW [49] e icoshift [50] já foi reportado na literatura, além dos resultados para modelos de regressão PLS.

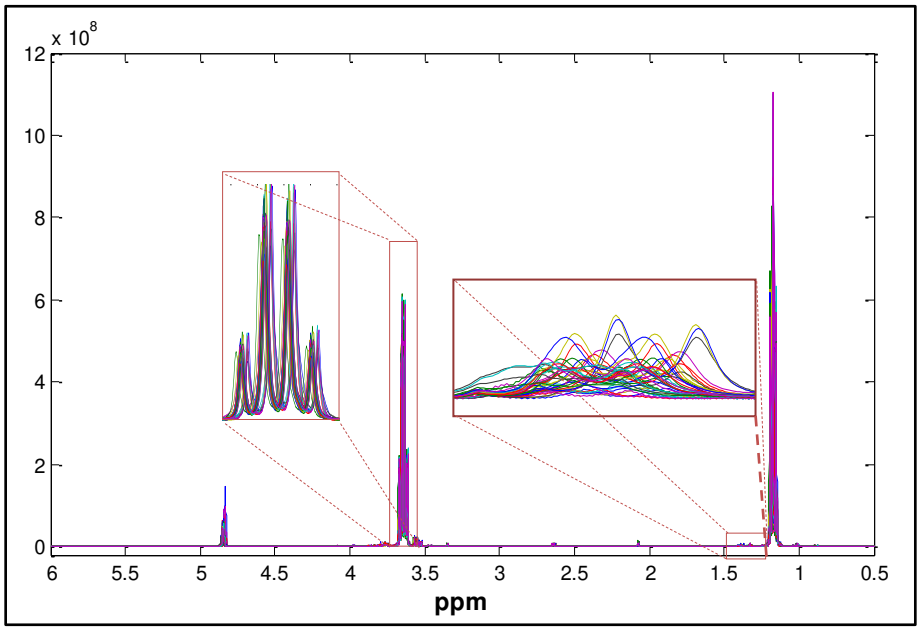

Figura 8: Espectros de RMN ^{1}H das amostras de vinho e regiões ampliadas relacionadas aos sinais do etanol (quarteto do grupo metileno) e ácido lático (dubleto do grupo metila terminal).

A Tabela 2 resume os resultados das regressões PLS (usando os dados centrados na média) para a calibração do conteúdo de ácido lático usando os dados originais e os dados pré-tratados pelos métodos *bucketing* convencional e OBA. Para comparação, os resultados utilizando as matrizes alinhadas por icoshift [50] e COW [49] também estão incluídos. Os modelos MLR para regiões pré-tratadas associadas com os deslocamentos químicos na região de 1,35 a 1,45 ppm correspondentes a dois *buckets* (#92 e #93 obtidos com o método *bucketing* conventional; e #91 e #92 obtidos com o método OBA) são também mostrados nesta tabela. Os dois *buckets* em cada situação apresentam baixa

correlação entre si (coeficientes de correlação 0,1363 e -0,0649, nos métodos *bucketing* convencional e OBA, respectivamente), assim não há redundância nos modelos MLR.

Baseando-se nos valores de R^2, os piores resultados foram aqueles relativos ao uso da faixa espectral inteira (menor valor de R^2). Possivelmente, a não linearidade introduzida pelos desalinhamentos impõe ao modelo PLS, o uso de um mais alto número de fatores (quatro variáveis latentes) para capturar a correlação entre os dados espectrais e o conteúdo de ácido lático. Além disso, pode ser visto que os modelos PLS construídos após as correções dadas por COW e icoshift são piores do que aquele para o dado original, tendo menores valores de R^2, com o mesmo número de variáveis latentes, que sugere que as correções não foram efetivas para resolver os desalinhamentos dos sinais do ácido lático. De fato, na publicação [49] relativa ao procedimento COW, os autores usaram uma ferramenta adicional baseada num procedimento em intervalos utilizando deslocamentos rígidos (coshift) para melhorar a calibração do ácido lático (veja na Tabela 2 a segunda linha referente aos resultados do método COW), uma vez que o método COW foi somente capaz de corrigir os desalinhamentos dos picos dominantes do etanol. Também na publicação [50] sobre o alinhamento por icoshift, os autores melhoraram a calibração para o ácido lático (veja na Tabela 2 a segunda linha referente aos resultados do método icoshift) usando intervalos personalizados, definidos através do conhecimento prévio das atribuições dos picos de RMN. Desta forma, os métodos COW e icoshift não funcionaram bem para o constituinte minoritário ácido lático, requerendo otimizações posteriores.

Tabela 2: Resultados dos modelos de regressão PLS e modelos MLR para o conteúdo de ácido lático (valores de referência, média =1,03 $g.L^{-1}$ e desvio padrão =0,51 $g.L^{-1}$).

Pré-tratamento	Região espectral (ppm)	#VL[a]	RMSECV[b] ($g.L^{-1}$)	R^{2c}
Nenhum (dado original)	0,5-6,0	4	0,369	0,48
	1,35-1,45	3	0,113	0,95
	1,35-1,45	2	0,136	0,93
icoshift[d]	0,5-6,0	4	0,400	0,39
	1,35-1,45	2	0,104	0,96
COW[e]	0,5-6,0	4	0,440	0,27
	1,3-1,6	3	0,100	0,96
Bucketing convencional	0,5-6,0 (*buckets*)	4	0,310	0,63
	1,35-1,45 (*buckets*)[f,g]	-	0,114	0,95
OBA	0,5-6,0 (*buckets*)	4	0,200	0,84
	1,35-1,45 (*buckets*)[f,h]	-	0,124	0,94

[a] #VL = número de variáveis latentes; [b] RMSECV = raiz quadrada do erro médio da validação cruzada (*Root Mean Squared Error of Cross Validation*); [c] R^2 = coeficiente de determinação; [d] da referência [50]; [e] da referência [49]; [f] modelos MLR; [g] *buckets* #92 e #93 (de 1,4608 até 1,4110 ppm = *bucket* #92 e de 1,4110 até 1,3611 ppm = *bucket* #93); [h] *buckets* #91 e #92 (de 1,4804 até 1,4261 ppm = *bucket* #91 e de 1,4261 até 1,3573 ppm = *bucket* #92); Os parâmetros são baseados na validação cruzada pela metodologia *leave-one-out*.

Os procedimentos de *bucketing* apresentaram desempenho superior para os modelos PLS, quando a faixa espectral inteira foi usada (Tabela 2), sem a necessidade de subsequentes manipulações. O tratamento pelo método *bucketing* convencional forneceu um modelo de regressão PLS com quatro variáveis latentes, R^2 = 0,63 e RMSECV = 0,310 $g.L^{-1}$ (Tabela 2), enquanto o método OBA alcançou R^2 = 0,84 com o

mesmo número de fatores (4 VLs) e com um menor erro na validação cruzada (RMSECV = 0,200 $g.L^{-1}$). Para evitar sobreajuste (*overfitting*), o número de variáveis latentes foi escolhido observando os gráficos de RMSECV *versus* o número de fatores, como mostrado na Figura 9. Como pode ser visto nesta figura, por exemplo, na curva relativa aos modelos PLS após o método OBA, aqueles com mais do que cinco variáveis latentes estão sobreajustados. Os melhores resultados após o pré-tratamento com OBA para o espectro inteiro podem estar associados à vantagem desta metodologia em concentrar os sinais em poucos *buckets*, evitando a divisão dos picos. De fato, o *bucketing* convencional é também capaz de alocar os sinais em poucos *buckets*, mas neste caso a divisão dos picos não é completamente evitada. Os resultados sugerem que para este conjunto de dados os procedimentos de *bucketing* são vantajosos, considerando os sérios desalinhamentos nos sinais do ácido lático.

Utilizando somente o sinal da metila terminal do ácido lático, o método icoshift mostrou o melhor modelo PLS com duas variáveis latentes (RMSECV = 0,104 $g.L^{-1}$ e R^2 = 0,96), enquanto o modelo PLS obtido após a correção fornecida pela abordagem COW gerou um modelo de regressão com estatísticas similares (RMSECV = 0,100 $g.L^{-1}$ e R^2 = 0,96), porém mais complexo com três variáveis latentes. Os modelos MLR obtidos a partir de cada procedimento de *bucketing* mostraram desempenho similar àquele obtido após a abordagem icoshift. Este fato constitui uma grande vantagem para as abordagens de *bucketing*, pois os modelos MLR não possuem viés (*unbiased models*), isto é, não são tendenciosos, além de serem mais simples do que os modelos PLS e eles não requerem a otimização do número de variáveis latentes.

O modelo MLR obtido após o uso do *bucketing* convencional (RMSECV = 0,114 $g.L^{-1}$ e R^2 = 0,95) foi ligeiramente melhor do que aquele obtido após OBA (RMSECV = 0,124 $g.L^{-1}$ e R^2 = 0,94), mas eles são ainda comparáveis, como verificado por um teste de variância (teste *F*) sobre os conteúdos de ácido lático preditos na validação cruzada para ambos os modelos MLR, a um nível de significância de α = 0,05, com número de

graus de liberdade no numerador e no denominador igual a 39 (valor de p = 0,9926). Adicionalmente, um teste de comparação de médias (teste t) foi realizado para os conteúdos de ácido lático, preditos pelos dois modelos MLR, onde foi encontrado que as médias não eram significativamente diferentes a um nível de significância de α = 0,05 (valor de p = 1,00).

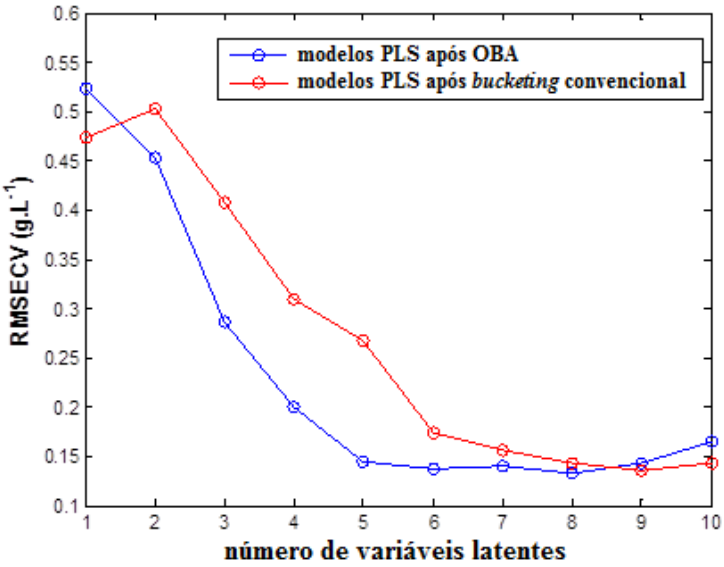

Figura 9: Valores de RMSECV *versus* o número de variáveis latentes obtidos na validação cruzada com a metodologia *leave-one-out* para os modelos PLS após OBA (linha azul) e *bucketing* convencional (linha vermelha).

Para a região específica estudada, os sinais das satélites dos ^{13}C do etanol podem ser encontrados (entre 1,30 e 1,35 ppm) muito próximos dos picos do ácido lático, que em parte pode dificultar a procura pelos mínimos locais e consequentemente afetar os modelos de regressão. Na verdade, os dois *buckets* (Tabela 2) usados para ambos os modelos MLR (após OBA e *bucketing* convencional) não cobrem exatamente a mesma região. Considerando este problema, uma possível solução é modificar os parâmetros no OBA (flexibilidade e largura inicial de *bucket*), porém isto não foi realizado no presente trabalho. Apesar disto, o método OBA proposto aqui funcionou muito bem para a

calibração do conteúdo de ácido lático a partir dos espectros de RMN ^{1}H das amostras de vinho.

2.4.2.2. Conjunto de dados de misturas de biodiesel-diesel

A Figura 10 mostra a comparação entre os espectros de RMN ^{1}H desalinhados e os mesmos espectros alinhados pelos métodos de *bucketing* convencional e OBA. Os espectros são referentes a cem amostras de misturas de biodiesel-diesel. É possível notar nas Figuras 10B e 10C o desempenho superior do método OBA para resolver o problema dos desalinhamentos (veja as regiões ampliadas). Os sinais de prótons das cadeias saturadas dos hidrocarbonetos no biodiesel e no diesel tendem a sofrer maiores deslocamentos nos picos devido às suas grandes liberdades conformacionais, fortemente dependentes da temperatura e que levam aos desalinhamentos vistos na Figura 10A. Um simples alinhamento dos espectros com respeito a um sinal de referência não pode corrigir tais deslocamentos.

As misturas são divididas em duas classes de acordo com a localização onde são comercializadas, como metropolitanas e não metropolitanas. As Figuras 11A, 11C e 11E comparam os gráficos de escores obtidos na PCA (primeira componente principal - PC1 *versus* segunda componente principal - PC2) para os espectros desalinhados centrados na média e para os espectros centrados na média, após alinhamento pelos métodos de *bucketing* convencional e OBA, respectivamente.

A análise dos gráficos de pesos nas Figuras 11B e 11D indicam que a distribuição das amostras é determinada pelos dois picos maiores, entre 0,7 e 1,4 ppm, sendo exatamente aqueles relacionados aos prótons mais afetados pelos desalinhamentos. Na Figura 11A, pode ser visto que PC1 captura a diferença entre os dois tipos de amostra, enquanto PC2 basicamente captura os desalinhamentos e a informação sobre duas amostras identificadas como “d” e “e” com valores de escores mais negativos. O

agrupamento na Figura 11A é similar àquele obtido na Figura 11C e é muito diferente daquele obtido utilizando o método OBA (Figura 11E), pois na Figura 11C, o alinhamento dos picos não é completamente realizado, como pode ser visto na região ampliada na Figura 10B e no gráfico de pesos na Figura 11D (similar àquele na Figura 11B).

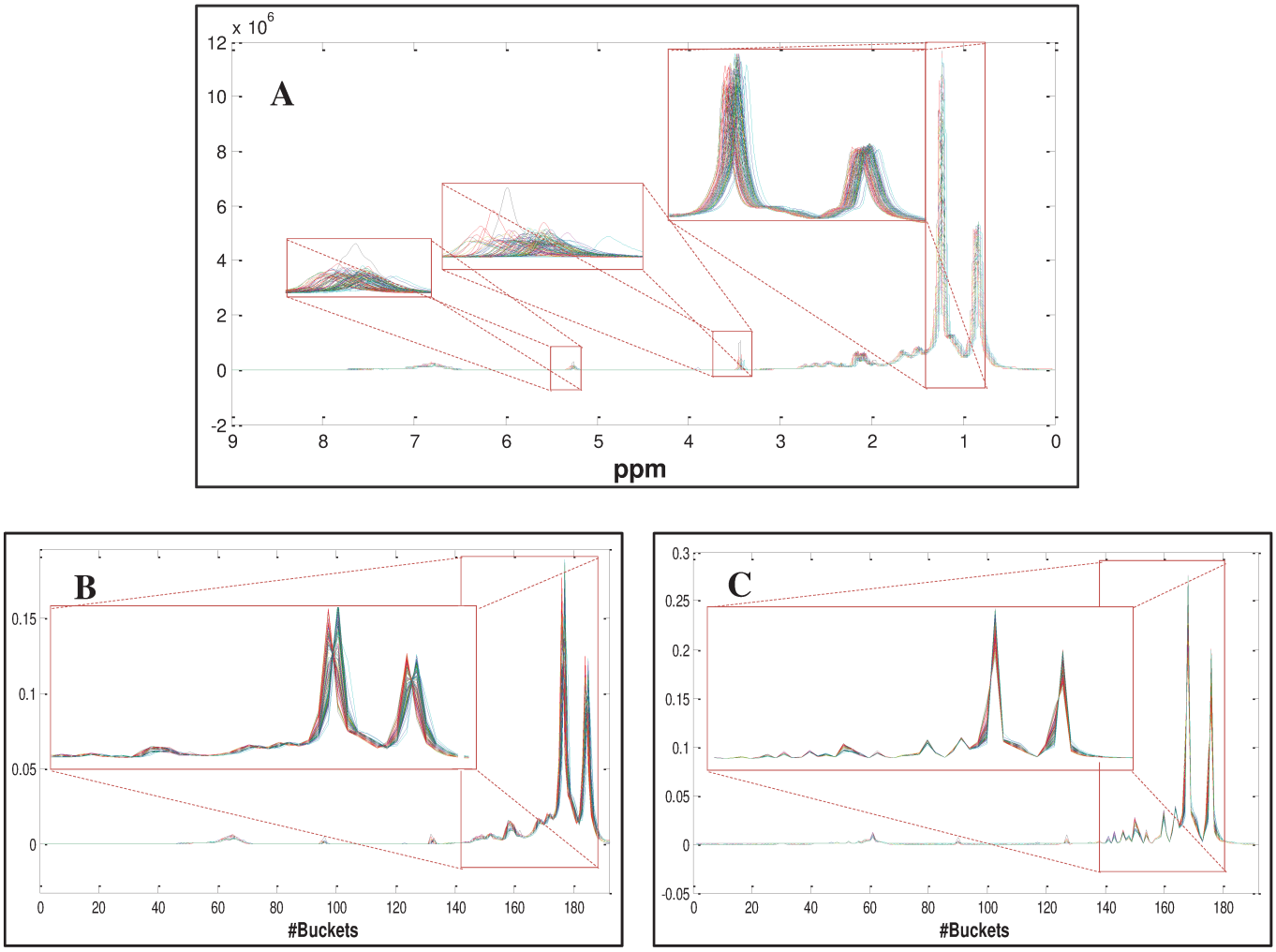

Figura 10: A: Os espectros originais com desalinhamentos evidenciados nas regiões ampliadas. Espectros pré-tratados por: B – *bucketing* convencional e C – OBA. As diferenças entre os valores nas ordenadas ocorrem porque os espectros pré-tratados são normalizadas para área unitária. Matrizes: B - **Z** (100 × 200); C - **Z** (100 × 191).

A partir de um conhecimento prévio obtido através de análises químicas padrões no campo dos combustíveis, para este conjunto de dados, todas as amostras identificadas

(a, b, c, d, e, f, g, h e i) são misturas não conformes, ou seja, fora das especificações. Portanto, como pode ser visto nas Figuras 11A e 11C, a variância associada aos desalinhamentos (capturada em PC2), neste caso, dificulta seriamente a identificação das amostras fora da especificação, não sendo interessante do ponto de vista da análise exploratória. Por outro lado, é claro na Figura 11E a evidência das amostras não conformes a partir dos seus próprios grupos com valores de escores extremos, tornando mais fácil a identificação e a interpretação. Após a correção fornecida pelo OBA, os escores obtidos descrevem 95,1% da variância total ao longo de PC1, praticamente toda a variância contida nas três primeiras componentes na PCA da matriz sem qualquer *bucketing* (PC1 74,6%, PC2 15,1% e PC3 5,7%) e 2,6% da variância total ao longo de PC2, que possui um gráfico de pesos (Figura 11F) completamente diferente daqueles obtidos nas outras análises. Neste exemplo, a complexidade desnecessária do perfil espectral fornecida pelos desalinhamentos foi corrigida apropriadamente.

Além da análise descrita acima, PCA foi realizada para os intervalos de 6,4 a 8,5 ppm e de 0,4 a 1,4 ppm, independentemente. O primeiro intervalo corresponde à região com sinais relativos aos prótons aromáticos que diferem daqueles em hidrocarbonetos, por sofrerem menos com os desalinhamentos (há um impedimento conformacional para os prótons em anéis aromáticos). De fato, os três agrupamentos obtidos para os dados centrados na média antes e após a aplicação de ambos os procedimentos de *bucketing* foram quase idênticos. Para o segundo intervalo (picos maiores), as observações foram muito similares àquelas citadas acima para a faixa espectral inteira. No capítulo 3, a seguir, as análises de componentes principais para estes intervalos, após a aplicação de OBA, serão mostradas e discutidas.

O método OBA também permitiu uma redução nos dados em cada espectro de RMN ^{1}H de 32768 pontos no domínio das frequências para 191 *buckets*. Este fato pode ser importante do ponto de vista computacional, uma vez que um reduzido número de variáveis pode diminuir significativamente o tempo computacional. Entretanto, vale

ressaltar que a redução é acompanhada por uma possível diminuição na resolução espectral, que pode levar a perdas de informação, especialmente quando diferenças sutis são expressas pelas amostras e alta resolução é requerida. Outras soluções para o alinhamento, tais como, os métodos DTW, COW e icoshift podem ser usados sem perda de resolução. Contudo, essas alternativas são teoricamente menos simples do que a abordagem de *bucketing* e comumente envolvem operações computacionais "caras" e uma maior *expertise* do usuário. Finalmente, é importante ressaltar que o método OBA fornece uma flexibilidade na definição dos parâmetros de entrada, que podem ser ajustados com o objetivo de evitar sérias perdas de resolução espectral.

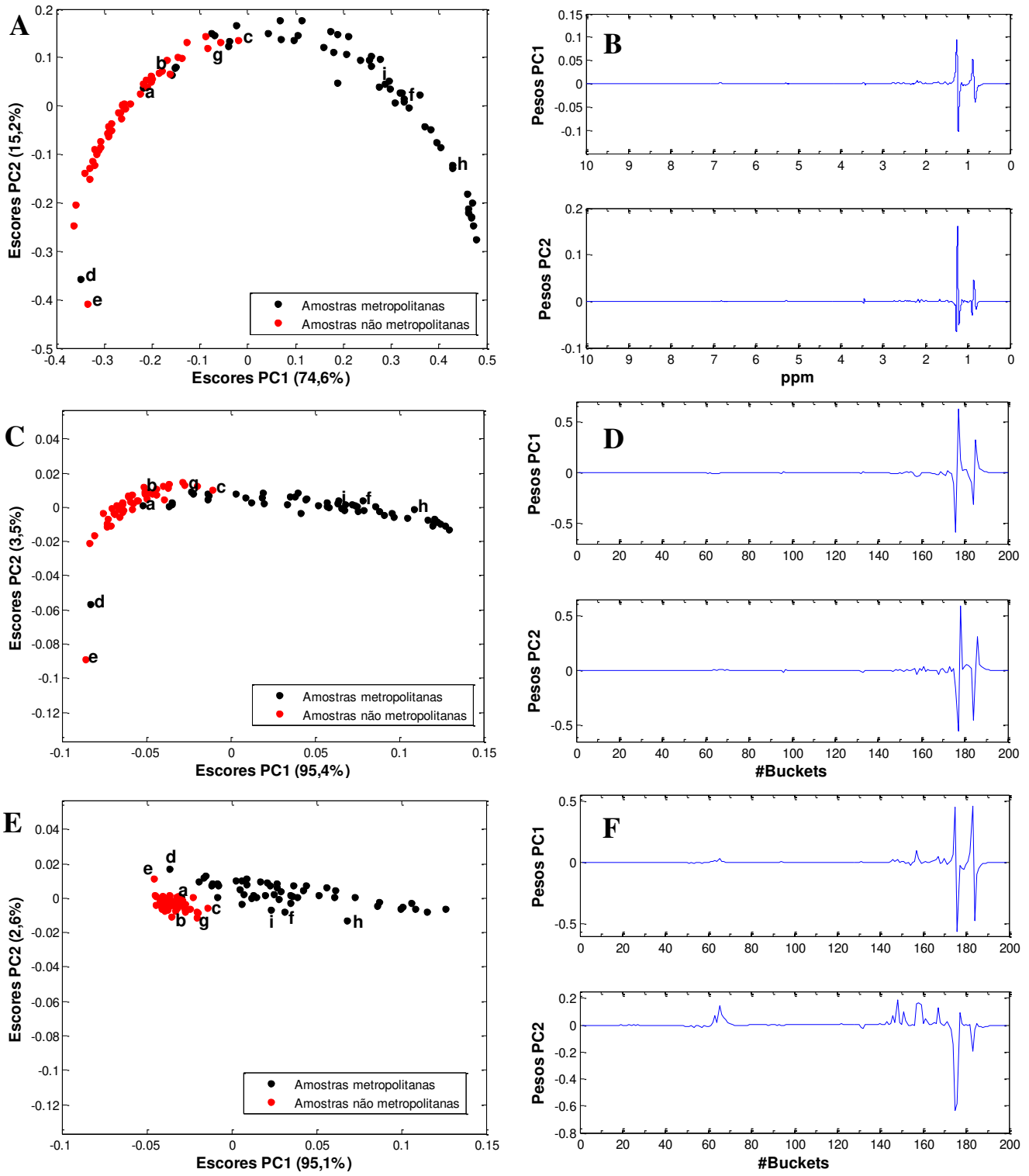

Figura 11: PCA: gráficos de (A) escores e (B) pesos para a matriz original; (C) escores e (D) pesos para a matriz após o *bucketing* convencional; e (E) escores e (F) pesos para a matriz após o método OBA. As variâncias explicadas são mostradas entre parênteses em cada componente principal.

2.4.2.3. Conjunto de dados de tumores cerebrais

As Figuras 12A e 12B apresentam os espectros de RMN ^{1}H para as duas classes de tumores cerebrais, não neuroglial (NN) e neuroglial em alto grau (Hg). A Figura 12A mostra os dados originais, onde é possível visualizar a pequena extensão dos desalinhamentos, através da região ampliada. Em contraste, a Figura 12B mostra os espectros corrigidos após o pré-tratamento com OBA, que resultou numa matriz de *buckets* **Z** com dimensões (29 × 1416). O pequeno valor de largura inicial de *buckets* de 0,002 ppm foi bem sucedido para corrigir os desalinhamentos, como pode ser observado na região ampliada da Figura 12B, mostrando que os picos nesta região tornaram-se mais nítidos e formatados.

Neste ponto, é importante citar o problema da correspondência comum para regiões complexas em espectros de RMN ^{1}H, onde devido a não homogeneidade do campo magnético durante a aquisição do espectro ou incompleta correção de fase no processamento dos dados adquiridos na análise, os formatos dos picos podem ser distorcidos do formato simétrico ideal e, além disso, as posições dos picos podem mudar (devido à temperatura, pH e força iônica) levando até mesmo a uma inversão na ordem dos sinais [58]. O método OBA aqui proposto não trata destas questões e a sua utilização deve ser feita em conjuntos de dados sem problemas extremos de correspondência. De fato, este é um ponto fraco inerente a todos os procedimentos de *bucketing*, sendo abordados por alguns trabalhos na literatura [59,60] propondo alternativas, como o uso da transformada Hough fuzzy generalizada (GFHT, do inglês *Generalized Fuzzy Hough Transform*) com o intuito de estabelecer objetivamente a correspondência verdadeira. Apesar disso, o *bucketing* ainda é amplamente utilizado, pois até agora nenhum método provou ser de fácil uso e bem sucedido na obtenção de bons resultados quanto ele [58]. Para o conjunto de dados de tumores cerebrais o problema da correspondência não ocorre, o que pode ser verificado na Figura 13, apresentando um gráfico de calor (*heat*

plot) criado para o conjunto de dados original (neste gráfico cada linha corresponde a um espectro cujas intensidades são codificadas pela cor), onde as amostras são ordenadas pelo valor de intensidade do sinal da creatina em 3,04 ppm (singleto), sendo a ordenação do menor (no topo) para o maior valor (em baixo). A ordenação não revela mudanças de ordem nos picos ao longo do eixo dos deslocamentos químicos, apenas o desalinhamento já observado. Vale ressaltar que nos estudos de caso anteriores não foi necessário checar o problema da correspondência, pois os espectros mostravam-se relativamente menos "povoados" do que os espectros estudados neste caso, e assim simples inspeções visuais indicaram a ausência do problema citado.

Com o objetivo de mostrar as vantagens do método OBA, modelos PLS-DA foram construídos para acessar a discriminação dos dois tipos de tumores. Os números de variáveis latentes usados em cada modelo foram definidos pela escolha daquele com o menor número de erros de classificação (NMC), evitando sobreajuste (*overfitting*) de modo similar àquele apresentado no estudo de caso do conjunto de dados de vinhos (Figura 9). Portanto, baseado neste diagnóstico estatístico, duas variáveis latentes foram determinadas na otimização do modelo PLS-DA relativo ao conjunto de dados centrado na média após o método OBA, onde o modelo obteve quatro erros na classificação entre os tipos de tumores. Para o modelo PLS-DA usando o conjunto de dados original centrado na média, quatro variáveis latentes com cinco erros na classificação foram selecionados. Ambos os modelos foram significantes a um nível de $\alpha = 0,05$ nos testes de permutação. As Figuras 14A e 14B mostram a distribuição de 10.000 testes de permutação para NMC dos modelos PLS-DA a partir do dado original e do pré-tratado. Observando a Figura 14, é visível que o número de permutações (10.000) usado foi suficiente para amostrar as caudas da distribuição, resultando em distribuições com formatos gaussianos. O modelo PLS-DA para a matriz após pré-tratamento com OBA obteve um valor de $p = 0,0032$ contra um valor de $p = 0,0005$ para o modelo PLS-DA relativo ao conjunto de dados original. Um valor de p menor do que o *threshold* (nível

ou limite) de significância $\alpha = 0,05$ indica que a hipótese nula H_0 (não há diferença entre as duas classes de tumores) deve ser rejeitada e, a este nível de significância, diferenças entre as classes são observadas.

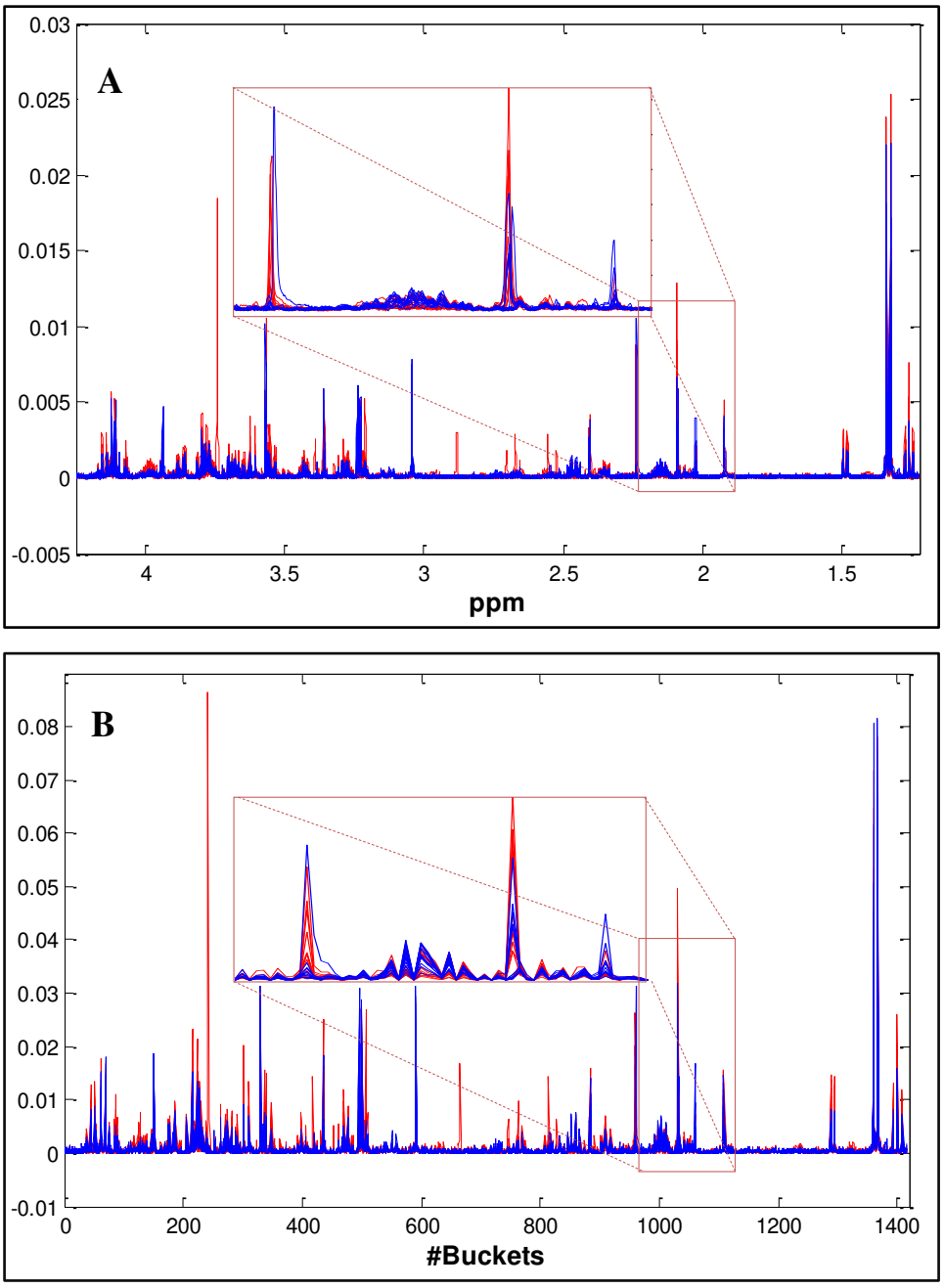

Figura 12: Conjunto de dados de tumores cerebrais. A: dados originais. B: após pré-tratamento com OBA (largura inicial de *buckets* de 0,002 ppm e flexibilidade de 50%). Espectros de RMN ^{1}H para tumores NN em vermelho e para tumores Hg em azul. Nas ordenadas temos as intensidades dos sinais. Os espectros encontram-se normalizados.

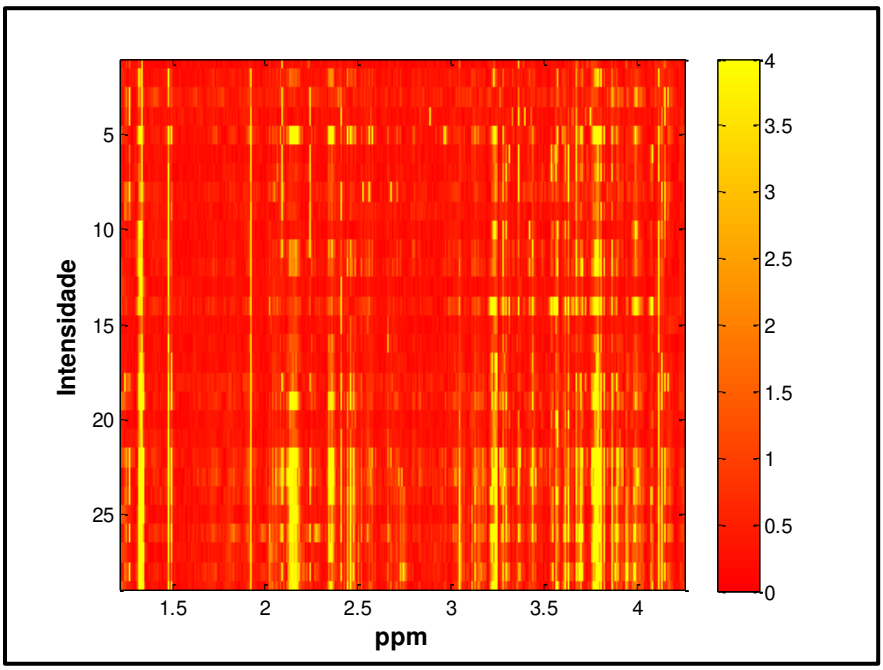

Figura 13: Espectros de RMN ^{1}H dos tumores cerebrais vistos num gráfico de calor (*heat plot*). As linhas representam as amostras e as colunas as intensidades em cada ppm. A intensidade é codificada de acordo com a barra de cor à direita. Os espectros encontram-se ordenados pelo valor de intensidade do sinal da creatina em 3,04 ppm (singleto), sendo a ordenação do menor (no topo) para o maior valor (em baixo).

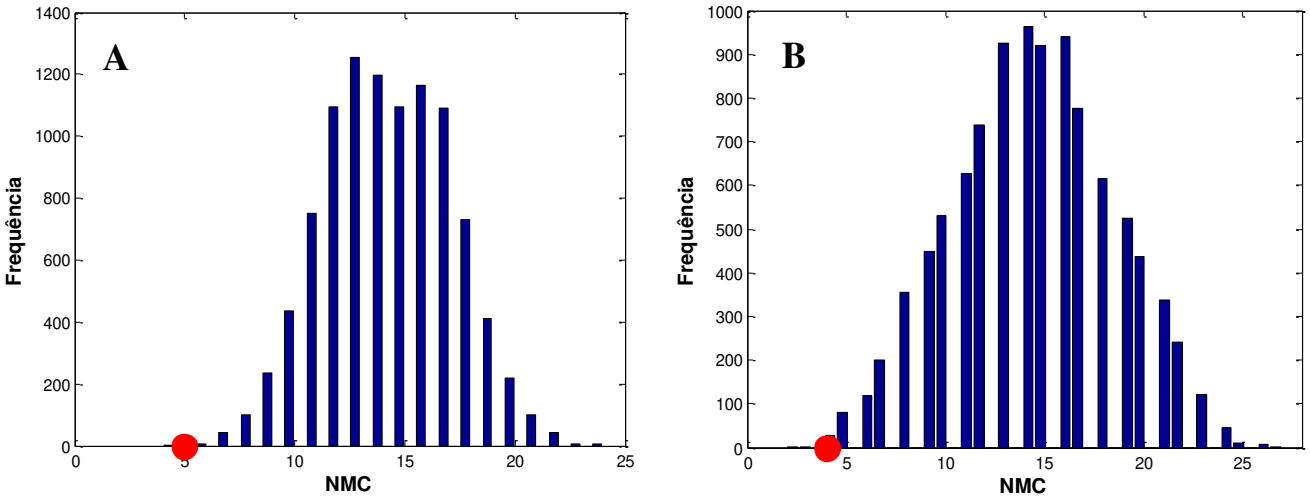

Figura 14: Distribuição de 10.000 testes de permutação para o NMC dos modelos PLS-DA do: A – conjunto de dados original e B – conjunto de dados após OBA. As bolas vermelhas indicam o NMC para cada modelo PLS-DA dos dados não permutados.

Para os dois modelos, as classificações foram significativas, entretanto, o modelo PLS-DA para os dados após o método OBA foi mais parcimonioso na determinação do subespaço de variáveis latentes que possibilita a discriminação (duas variáveis latentes contra quatro para o outro modelo), além de obter uma melhoria no NMC.

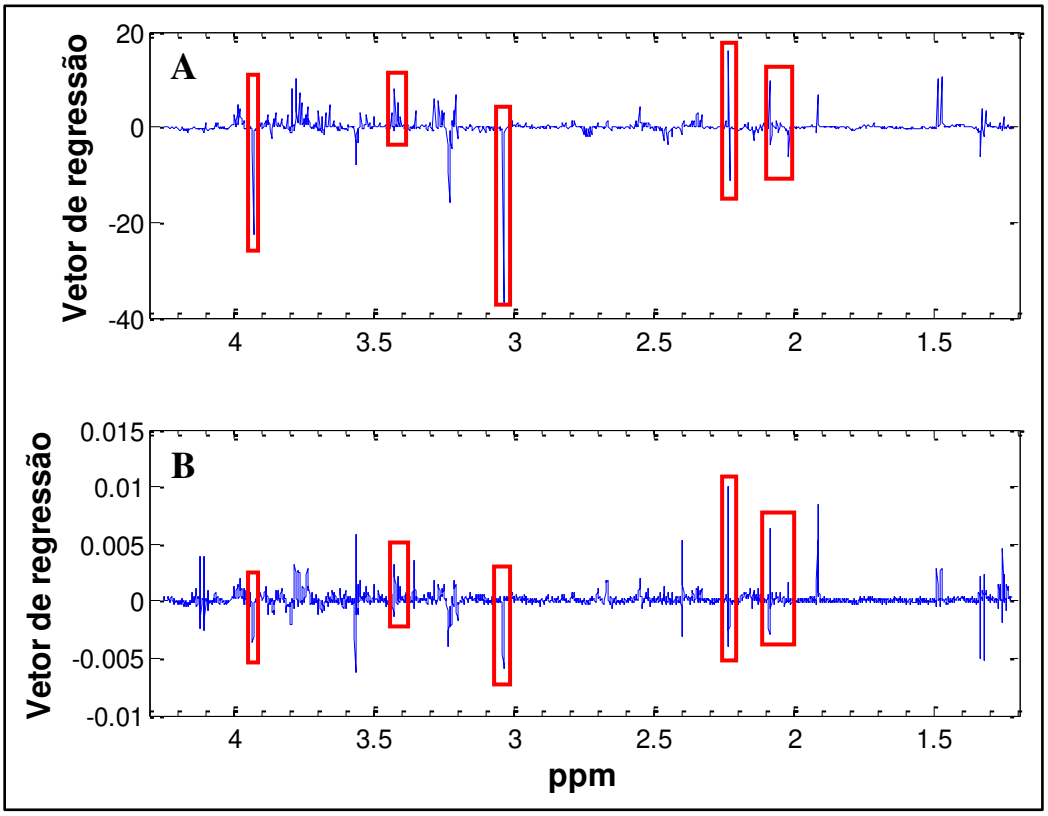

Figura 15: Vetores de regressão, representados nos eixos dos deslocamentos químicos, para A: modelo PLS-DA com duas variáveis latentes a partir dos dados após OBA e B: modelo PLS-DA com quatro variáveis latentes a partir dos dados originais.

As Figuras 15A e 15B mostram os vetores de regressão obtidos para os modelos PLS-DA a partir dos dados após o método OBA e sem a aplicação deste, respectivamente. Nas figuras os retângulos vermelhos marcam os sinais mais discriminantes entre os dois tipos de tumores, conforme apontado pela literatura [33]. Analisando o vetor de regressão na Figura 15A, pode ser visto, comparando com o vetor de regressão na Figura 15B, que os sinais importantes tornam-se mais evidentes e com importância relativa maior após a redução na complexidade do dado devido ao pré-

tratamento pelo método OBA, sendo isto importante no contexto da busca por biomarcadores.

Finalmente, além da superioridade mostrada devido ao "alinhamento" fornecido pelo método OBA, há a vantagem pela redução do número de variáveis, que pode ser muito importante do ponto de vista computacional, especialmente quando são tratados grandes conjuntos de dados. Este exemplo mostra a grande aplicabilidade do novo algoritmo e sua flexibilidade para resolver questões em um conjunto de dados complexo.

2.5. Conclusões

Foi mostrado que o método OBA tem um desempenho superior, quando comparado ao *bucketing* convencional, largamente usado na literatura. Para o conjunto de dados de vinhos, os resultados demonstraram que a estratégia do método OBA pode ser útil para a construção de modelos menos complexos (modelos MLR) com boas capacidades preditivas, mesmo comparados a modelos PLS obtidos quando sofisticados métodos de alinhamento são usados. Para o conjunto de dados de misturas de biodiesel-diesel, foi mostrado o bom desempenho do método OBA como auxiliar na análise exploratória, que pode ser de grande significância para propósitos de reconhecimentos de padrões. O principal ponto reside na melhoria (aumento) da variância explicada pelas componentes principais com consequente aumento na interpretabilidade. Neste exemplo, o método OBA forneceu bons resultados, mesmo num conjunto de dados com uma grande extensão de desalinhamentos. No conjunto de dados de tumores cerebrais, o pré-tratamento por OBA permitiu a obtenção de um modelo PLS-DA significativo para a discriminação dos tumores com um menor número de erros na classificação, corrigindo um pequeno número de desalinhamentos, especialmente importantes para os sinais mais discriminatórios. O algoritmo proposto neste trabalho é fácil de usar, sendo necessário para os usuários somente o conhecimento da extensão dos desalinhamentos na linha de

base para especificar um tamanho inicial adequado de *bucket* e flexibilidade. Este ponto é muito importante, pois a metodologia de *bucketing* fornece uma diminuição na resolução espectral, que pode ser minimizada pelo uso de parâmetros de entrada adequados. Em nossas aplicações, uma flexibilidade de 50% foi adequada, mas não é possível generalizar a respeito disso, uma vez que diferentes conjuntos de dados podem requerer diferentes valores. O algoritmo consiste de uma rotina escrita em código Matlab e encontra-se disponível para *download* gratuito em http://lqta.iqm.unicamp.br.

Capítulo 3: Metodologia para controle de qualidade de misturas de biodiesel-diesel por análise de componentes principais em intervalos (iPCA) e espectroscopia de ressonância magnética nuclear de próton e de carbono (RMN ^{1}H e ^{13}C)

3.1. Introdução

Nos últimos anos as pesquisas com biocombustíveis têm recebido muita atenção da comunidade científica, devido entre outros fatores, à crescente preocupação ambiental, ao esgotamento das reservas de combustíveis fósseis e ao aumento da demanda mundial de energia. Assim, alternativas para a substituição dos combustíveis fósseis, seja total ou parcial, têm sido buscadas através da exploração de fontes renováveis de energia. Neste contexto, o biodiesel tem se mostrado uma boa alternativa para a substituição do diesel combustível, pois apresenta diversas vantagens como: características físico-químicas similares ao diesel; é virtualmente livre de enxofre, biodegradável e não tóxico; possui alto número de cetano (parâmetro relacionado à qualidade de ignição), alto ponto de fulgor (relativo à capacidade do combustível em formar uma mistura inflamável com o ar) e alta lubricidade; não requer modificações na infraestrutura já existente para o diesel, no que diz respeito à estocagem e distribuição [61-68].

Quimicamente, o biodiesel é uma mistura de monoalquil ésteres de ácidos graxos que são obtidos através da reação de transesterificação, catalisada por base, de gorduras animais ou óleos vegetais com um álcool de cadeia curta (e.g., metanol ou etanol), tendo como subproduto o glicerol. Existem diversas matérias-primas potenciais para a produção do biodiesel, porém os óleos vegetais são aquelas que mais se destacam [69-72].

O Brasil é um país que, por seu clima tropical e subtropical aliados à sua extensa área geográfica, favorece uma grande diversidade de matérias-primas para a produção de biodiesel. De fato, o país é considerado mundialmente como tendo uma posição estratégica para a sustentabilidade do mercado do biodiesel e desde 2005 possui legislação (Lei n° 11097, de 13 de janeiro de 2005) que regulamenta as condições de mercado e a produção, propiciando a introdução do biodiesel em sua matriz energética. Atualmente no Brasil, todo o diesel comercializado obrigatoriamente deve possuir 5% de biodiesel na sua composição [73].

O diesel mineral consiste numa mistura bastante complexa, composta principalmente por hidrocarbonetos aromáticos (em menor extensão) e alifáticos (*n*-parafinas, naftenos, *iso*-parafinas, e em menor quantidade olefinas, que podem estar presentes devido a misturas com derivados de *cracking* do petróleo) com ampla distribuição em termos de tamanho de cadeia carbônica e ramificações. Além disso, o diesel possui compostos aromáticos mono e poli nucleados que têm átomos de enxofre na estrutura, tais como, tiofenos (mononuclear), benzotiofenos (polinuclear), dibenzotiofenos (polinuclear) e os seus derivados [74,75].

No Brasil o diesel comercializado nas grandes cidades (regiões metropolitanas) é diferente daquele comercializado nas demais regiões, principalmente no que diz respeito ao teor de enxofre, o que por sua vez envolve questões ambientais relativas à qualidade do ar nos grandes centros. Para a diferenciação imediata destes tipos de diesel, estes são comercializados com diferentes colorações, sendo adicionado àqueles das regiões não metropolitanas, corantes que dão uma tonalidade vermelha, enquanto os outros preservam a cor típica do diesel que varia entre amarelada e alaranjada [76].

Na literatura, no estudo das propriedades do biodiesel puro e de suas misturas com o óleo diesel, existe um grande número de trabalhos aplicando diversas técnicas analíticas, tais como, cromatografia gasosa [77-80], espectroscopia no infravermelho médio e próximo [68,81-83], ressonância magnética nuclear [16,71,77-82,84-86]. Nestas

áreas, em geral, as pesquisas visam proporcionar alternativas instrumentais às análises comuns dos biocombustíveis, tendo em vista que estas normalmente são realizadas através de técnicas laboriosas, demoradas e que consomem uma relativa maior quantidade de reagentes.

3.2. Proposta geral do trabalho

Dentre os trabalhos usando a espectroscopia de ressonância magnética nuclear (RMN) para estudo do biodiesel, poucos são aqueles que usam está técnica combinada com ferramentas quimiométricas [16,71]. As vantagens da RMN já foram comentadas no capítulo anterior e podem ser exploradas juntamente com um tratamento de dados apropriado para a análise de misturas de biodiesel-diesel. Com este espírito, neste trabalho exploram-se os espectros de RMN ^{1}H e ^{13}C destas amostras, adquiridos numa condição padrão, para a detecção daquelas fora das especificações, ou seja, as amostras discrepantes. Os dados de análises químicas relativos à determinação de alguns parâmetros próprios do campo de biocombustíveis são conhecidos previamente para as amostras e assim este trabalho visa verificar se o padrão espectral responde da mesma forma que estas análises.

Para o uso dos espectros de RMN ^{1}H a questão dos desalinhamentos é importante para o tipo de amostra aqui estudado, no entanto, a solução para isto já foi discutida no capítulo anterior.

No caso dos espectros de RMN ^{13}C, a principal preocupação se refere ao ruído espectral que em geral representa uma parte significativa dos dados, pois como o átomo de ^{13}C é pouco abundante (cerca de 1,1% do total dos átomos de C) a relação sinal/ruído tende a ser menor [87]. Na prática uma solução para o aumento dessa relação é obtida acumulando vários FIDs (entre vários pulsos e entre tempos necessários para a relaxação dos núcleos), porém para a aplicação aqui estudada, pensada para o uso de análises de

rotina, esta solução não é atrativa, pois gera um maior tempo de análise. Por outro lado, usando poucas acumulações (*scans*) na aquisição dos espectros de RMN ^{13}C, mesmo aplicando transformações que processam (operações para obtenção dos espectros que são realizadas após a aquisição dos dados) os FIDs e que podem aumentar a relação sinal/ruído, a quantidade de ruído é ainda considerável. Felizmente, existem ferramentas matemáticas que podem ser usadas após o processamento, podendo auxiliar na redução do ruído espectral.

Existem diferentes modos de reduzir o ruído de espectros, sendo os mais comuns, os filtros de Fourier [88] e *wavelet* [43,89-92]. Estas abordagens decompõem os sinais como uma combinação linear de um conjunto particular de funções de base, reconstruindo-os depois de uma filtragem apropriada. O filtro *wavelet* tem algumas vantagens sobre a transformada de Fourier, pois suas funções de base são melhores localizadas no domínio do tempo e podem descrever fenômenos não estacionários [43,91,92]. Por esta razão a transformada *wavelet* é muita atrativa para a redução de ruído (*denoising*) dos espectros de RMN ^{13}C.

O processamento dos FIDs, citado acima, é muito importante para uma posterior análise quimiométrica, pois o tratamento de dados, em geral, é capaz de distinguir as amostras processadas diferentemente. Assim, todos os passos envolvidos no processamento precisam seguir um mesmo protocolo para todas as amostras, de modo que não haja influência nas inferências estatísticas. Tal processamento compreende os passos de tratamento do FID (multiplicação por funções de apodização, preenchimento do sinal com zeros, truncagem, entre outras); aplicação da transformada de Fourier (conversão do sinal adquirido no domínio do tempo para um sinal no domínio da frequência); ajuste de fases; e correção da linha de base [87].

Atualmente, vários programas comerciais incorporam múltiplos métodos para o processamento automático de espectros de RMN. Entretanto, no caso de misturas complexas, como as amostras aqui estudadas, que resultam em espectros de RMN

complexos, com regiões sobrepostas, o processamento, como o ajuste de fase, torna-se mais difícil, pois os métodos propostos são eficientes apenas para sinais bem resolvidos [42]. Portanto, neste trabalho, todos os espectros de RMN ^{1}H e ^{13}C tiveram suas fases cuidadosamente corrigidas manualmente. Já as linhas de base foram corrigidas automaticamente usando uma função linear.

A PCA é *de facto* o método de análise exploratória em quimiometria, contando com um número enorme de publicações neste sentido. Porém, nem sempre a PCA evidencia facilmente os agrupamentos ou respostas esperados, pois isto também dependente do pré-processamento dos dados, da intensidade relativa de sinais importantes, da seleção de variáveis realizada, entre outros. Por exemplo, se num conjunto de dados um sinal é importante na discriminação de um determinado tipo de amostra com outras de tipo diferente, mas o mesmo é pouco intenso em relação aos demais e é praticamente "mascarado", dificilmente a PCA irá detectá-lo com os dados centrados na média. Por outro lado, se os dados forem autoescalados, há uma chance deste sinal ser mostrado, uma vez que após o autoescalamento a matriz resultante apresenta a mesma importância em todas as suas variáveis (as variâncias se tornam iguais e com valor um). Infelizmente, o autoescalamento também irá enfatizar a importância do ruído espectral e a resposta da PCA cai num novo problema.

Neste trabalho, a análise de componentes principais em intervalos (iPCA) que corresponde a uma extensão da PCA comum, onde intervalos no conjunto de dados são usados, ao invés do conjunto todo, é usada com o objetivo de melhorar a interpretação dos resultados, explorando modelos para variações locais e diminuindo o ocultamento dos sinais menos intensos.

3.3. Objetivos

3.3.1. Geral

O objetivo geral deste trabalho é estudar uma metodologia que aplicada a amostras de misturas de biodiesel-diesel visa auxiliar no controle de qualidade das mesmas, valendo-se da exploração de espectros de RMN ^{1}H e ^{13}C. Também se deseja verificar a relação entre as respostas obtidas a partir dos padrões contidos nos espectros de RMN com os resultados de análises químicas típicas do campo de biocombustíveis.

3.3.2. Específicos

Especificamente, objetiva-se:

- Aplicar a iPCA para evidenciar intervalos nos dados que fornecem diferentes observações acerca das amostras não conformes e resultem num ganho de interpretabilidade;
- Verificar se o estudo dos espectros de RMN ^{13}C promove algum ganho de informação em relação ao estudo dos espectros de RMN ^{1}H. Neste contexto, visa-se também avaliar a viabilidade do uso da redução de ruído (*denoising*) dos espectros de RMN ^{13}C e do curto tempo de aquisição destes espectros;
- Observar se existem problemas na análise devido ao não uso da trava de frequência (*locking* de campo magnético) na aquisição dos espectros de RMN.

3.4. Parte Experimental

3.4.1. Amostras

A descrição das amostras estudadas aqui se encontram no item 2.4.1.2 do capítulo anterior. Alguns resultados relativos a parâmetros de qualidade avaliados para as amostras de biocombustíveis foram adquiridos juntos à Central Analítica do Instituto de Química da Unicamp e são apresentados na Tabela 3. Nesta tabela são destacadas com fundo cinza as amostras que foram detectadas como não conformes pela metodologia aqui apresentada, marcadas com asteriscos as amostras detectadas como não conformes pelos parâmetros físico-químicos da própria tabela e em fundo azul as amostras detectadas em ambos.

3.4.2. Aquisição dos espectros de RMN ^{1}H e ^{13}C

A descrição relativa à aquisição dos espectros de RMN ^{1}H das amostras de misturas de biodiesel-diesel está presente no item 2.4.1.2 do capítulo anterior. Para os espectros de RMN ^{13}C, as análises de RMN foram realizadas em um espectrômetro Bruker Avance DRX400 operando a uma frequência de 100,62 MHz para ^{13}C à temperatura ambiente, usando 550 μL de amostra pura, com uma sonda Bruker BBO, sem rotação, usando uma sequência de pulso padrão para $^{13}C\{^{1}H\}$. A homogeneidade do campo foi obtida pela inspeção do formato de uma linha espectral de uma amostra padrão (0,3% de clorofórmio em acetona-d6). Esta condição de campo foi usada para todas as amostras durante todas as análises. Todos os espectros foram adquiridos com 32768 pontos (32k) no domínio do tempo, 260 ppm de largura de janela espectral e 16 varreduras (*scans*). Os FIDs foram processados com o *software* TOPSPIN 2.1 com 65536 pontos (64k), multiplicados por uma função de janela exponencial com largura de

linha constante de 10 Hz e transformada de Fourier normal. Os espectros finais tiveram suas fases ajustadas, um a um, por inspeção direta. A correção da linha de base foi feita usando uma função linear automática. Todos os espectros foram referenciados usando uma posição de campo digital obtida usando tetrametilsilano (TMS) em acetona a 0 ppm.

Neste trabalho, até onde se sabe, pela primeira vez adquiriram-se espectros de RMN de misturas de biodiesel-diesel puras sem trava de frequência. Isto foi feito para minimizar o manuseio no preparo das amostras e eliminar o consumo de solvente deuterado (o sinal de ressonância do deutério é utilizado para a trava de frequência, ou seja, a correção de pequenas variações no campo magnético).

3.4.3. *Bucketing* otimizado

A descrição do algoritmo de *bucketing* otimizado (OBA) foi mostrada no capítulo 2, sendo a matriz com os espectros de RMN ^{1}H resultante do tratamento do conjunto de dados de misturas de biodiesel-diesel, estudada aqui. A relação entre cada *bucket* e a faixa em ppm nos espectros de RMN ^{1}H é mostrada na Tabela 4.

3.4.4. Redução do ruído dos espectros de RMN ^{13}C

Neste trabalho, os espectros de RMN ^{13}C foram adquiridos com um número reduzido de varreduras, com o objetivo de avaliar a possibilidade da utilização desta técnica para análises de rotina, onde o tempo é um fator importante. No entanto, isso favorece a aquisição de espectros ruidosos. Desse modo, avaliou-se a viabilidade do uso do método de análise de componentes principais em múltiplas escalas (MSPCA, do inglês *Multiscale Principal Component Analysis*) que se baseia na transformada *wavelet* para permitir uma redução no ruído espectral. Aspectos teóricos sobre a transformada

wavelet, bem como sobre a etapa de redução no ruído usando MSPCA serão mostrados a seguir. Todos os cálculos foram feitos utilizando as rotinas da Wavelet toolbox versão 4.3 para Matlab (The MathWorks, Natick, MA, USA).

3.4.4.1. Transformada *Wavelet* Discreta

A transformada *wavelet* (TW) surgiu como uma alternativa para contornar alguns problemas existentes no uso do filtro de Fourier e do filtro de Fourier em janelas, e tem se mostrado, desde o seu principal desenvolvimento, em meados da década de 80, uma excelente abordagem para filtragem ou compressão de dados. Enquanto a transformada de Fourier expressa um sinal como uma combinação de funções senos e cossenos que se estendem de menos infinito até mais infinito e são simétricas, a TW utiliza funções assimétricas e com energia concentrada para decompor o sinal. Além disso, em química, uma grande vantagem da TW frente a outros filtros mais comuns, tais como, filtros de média, média móvel e Savitsky-Golay, é a sua habilidade em tratar sinais estreitos sem distorcê-los [93-95].

A transformada *wavelet* discreta (DWT, do inglês *Discrete Wavelet Transform*) consiste na projeção de algum sinal discreto sobre um conjunto de funções *wavelet* que derivam de uma função *wavelet* mãe e que compõem uma família de funções. A forma geral dessas funções encontra-se na Equação 17 abaixo:

$$\psi_{m,k}(t) = 2^{-m/2}\psi(2^{-m}t - k) \qquad \text{Equação 17}$$

Onde $\psi(t)$ representa a *wavelet* mãe e m e k são os parâmetros diádicos de dilatação e translação, respectivamente. O parâmetro de dilatação determina a localização no domínio da frequência (ou escala), enquanto o parâmetro de translação determina a localização da função no domínio do tempo [90].

Tabela 3: Parâmetros analisados para as amostras de biodiesel-diesel. As amostras detectadas somente pela metodologia apresentada são mostradas com fundo cinza; aquelas reprovadas pelos parâmetros físico-químicos são marcadas com asteriscos; e aquelas reprovadas por ambos são mostradas com fundo azul.

Amostras		Teor de Biodiesel[a] (%)	Massa específica[b] (Kg/m^3)	T50%[c] (°C)	T85%[d] (°C)	Índice de cetano[e]	Teor de Enxofre[f] (mg/Kg)	Ponto de Fulgor[g] (°C)
1	metrop.	3,60	850,0	256,10	348,30	42,10	206	40,50
2 = a*	metrop.	3,30	858,0	294,90	351,10	46,30	251	52,00
3	metrop.	3,50	847,0	250,80	343,40	42,40	219	41,50
4	metrop.	4,30	850,0	258,80	348,90	42,40	410	41,00
5	metrop.	4,10	850,0	260,00	349,20	42,90	280	40,50
6	metrop.	4,20	848,0	254,90	346,80	42,70	286	40,50
7 = n*	metrop.	3,30	846,0	253,20	345,40	42,80	263	42,00
8*	metrop.	3,40	850,0	254,50	347,80	41,90	187	44,00
9*	metrop.	3,40	851,0	260,20	349,00	42,50	169	44,00
10*	metrop.	3,30	850,0	257,80	348,00	42,40	166	35,50
11 = gg	metrop.	4,50	847,0	254,00	349,80	42,80	161	39,50
12	metrop.	4,20	852,0	259,80	348,50	42,10	400	45,00
13	metrop.	4,00	850,0	257,10	349,00	42,30	233	39,00
14 = rr*	metrop.	3,20	850,0	264,70	350,00	43,50	259	45,00
15	metrop.	3,60	849,0	257,00	347,80	42,60	143	38,50
16 = m*	metrop.	3,50	850,0	250,50	347,10	41,50	148	39,50
17*	metrop.	3,40	850,0	252,60	345,40	41,60	163	40,50
18 = pp*	metrop.	3,40	850,0	254,90	346,80	41,80	187	38,50
19 = t*	metrop.	4,10	855,0	267,40	354,20	42,40	148	25,00
20	metrop.	4,40	848,0	256,50	347,80	42,90	276	40,00
21 = l	não metrop.	4,20	856,0	292,60	354,50	46,00	1060	39,50
22	metrop.	3,60	849,0	260,70	348,70	43,10	300	45,00
23	metrop.	4,20	853,0	264,40	350,20	42,50	379	43,50
24	metrop.	4,40	848,0	256,50	347,80	42,90	276	40,00
25	metrop.	4,40	851,0	270,00	351,40	44,00	274	44,00
26	metrop.	4,30	846,0	253,90	347,50	43,10	256	40,50
27 = r	metrop.	4,30	845,0	253,90	346,20	43,30	243	41,00
28 = p*	metrop.	3,00	849,0	267,10	356,90	43,80	209	38,00
29 = u*	não metrop.	2,50	855,0	286,80	361,00	44,60	1600	41,50
30 = ss*	não metrop.	3,40	855,0	287,50	359,70	45,10	1700	43,00
31 = jj*	metrop.	3,40	852,0	261,30	351,60	42,40	400	45,00
32 = tt	metrop.	3,70	853,0	260,90	347,80	42,00	400	46,00
33 = kk*	não metrop.	3,20	854,0	288,20	359,20	45,30	1700	45,00
34*	não metrop.	2,50	855,0	285,30	356,40	44,30	1500	40,41

Amostras		Teor de Biodiesel[a] (%)	Massa específica[b] (Kg/m^3)	T50%[c] (°C)	T85%[d] (°C)	Índice de cetano[e]	Teor de Enxofre[f] (mg/Kg)	Ponto de Fulgor[g] (°C)
35 = z*	não metrop.	3,40	854,0	288,30	359,60	45,40	1700	44,00
36 = j	não metrop.	4,00	846,0	279,90	337,80	44,90	1000	50,00
37*	não metrop.	3,20	854,0	282,40	353,60	44,60	1800	43,00
38*	não metrop.	2,50	855,0	286,20	358,50	44,40	1800	40,41
39 = w*	metrop.	3,40	852,0	260,20	351,70	42,00	400	45,50
40 = b*	não metrop.	3,20	854,0	290,00	365,10	45,40	1600	40,41
41	não metrop.	3,80	855,0	286,60	357,90	45,00	1700	43,00
42	metrop.	4,30	854,0	262,60	349,20	42,00	400	46,50
43 = c	não metrop.	4,50	854,0	285,80	356,60	45,00	1700	42,00
44 = bb*	não metrop.	3,20	855,0	285,80	358,00	44,80	1700	43,50
45 = q*	não metrop.	2,50	855,0	285,30	356,40	44,30	1500	40,41
46*	não metrop.	3,40	855,0	289,30	359,90	45,20	1700	44,00
47	não metrop.	4,10	855,0	283,10	353,50	44,10	1600	42,00
48	metrop.	4,10	852,0	260,90	348,80	42,20	300	44,00
49	não metrop.	3,80	855,0	286,00	357,30	44,90	1700	40,00
50	metrop.	4,30	851,0	257,00	349,60	42,10	381	41,00
51	metrop.	4,00	852,0	261,20	350,30	42,20	300	42,00
52	metrop.	4,40	846,0	253,50	345,40	43,20	300	38,00
53*	não metrop.	3,40	856,0	289,60	364,50	44,80	1500	44,00
54 = s*	não metrop.	3,20	856,0	283,70	348,90	44,50	1400	44,50
55	não metrop.	4,20	855,0	285,00	355,60	44,50	1700	43,50
56 = x*	não metrop.	3,40	853,0	281,10	354,10	44,50	1600	40,50
57 = cc*	não metrop.	3,40	854,0	285,00	355,80	44,70	1600	43,50
58	metrop.	4,20	852,0	262,20	352,20	42,30	400	45,00
59	metrop.	4,40	849,0	259,50	350,70	43,20	306	42,00
60 = y*	não metrop.	3,10	855,0	287,20	361,30	45,00	1800	43,00
61 = ll*	não metrop.	3,30	854,0	280,90	350,20	44,40	1500	44,00
62 = dd	metrop.	4,00	848,0	264,40	355,10	43,70	400	41,00
63 = qq*	não metrop.	3,20	855,0	289,00	360,90	45,20	1500	44,00
64	não metrop.	3,60	855,0	286,20	356,60	44,70	1700	44,00
65*	não metrop.	3,30	855,0	285,20	353,50	44,20	1700	43,50
66	não metrop.	4,20	855,0	289,20	359,00	45,20	1600	43,00
67*	não metrop.	3,40	853,0	281,80	355,40	44,70	1700	41,50
68 = ff	metrop.	4,30	849,0	259,60	352,30	43,10	300	43,00
69	metrop.	4,20	852,0	261,50	351,90	42,30	400	45,00
70	metrop.	3,60	850,0	257,70	349,10	42,30	195	40,00
71 = ii*	metrop.	3,50	851,0	256,10	347,40	41,90	210	40,50
72*	metrop.	3,10	850,0	253,30	345,90	41,80	203	39,00
73*	metrop.	3,40	848,0	253,60	344,20	42,30	214	40,00

Amostras		Teor de Biodiesel[a] (%)	Massa específica[b] (Kg/m^3)	T50%[c] (°C)	T85%[d] (°C)	Índice de cetano[e]	Teor de Enxofre[f] (mg/Kg)	Ponto de Fulgor[g] (°C)
74	não metrop.	4,40	855,0	284,50	354,90	44,40	1600	40,50
75*	metrop.	3,10	850,0	264,90	354,30	43,40	139	41,50
76*	não metrop.	3,20	855,0	286,30	354,80	44,20	1500	41,00
77*	metrop.	4,10	850,0	262,60	353,50	43,30	147	33,00
78 = d*	metrop.	2,80	859,0	289,70	347,90	45,50	224	49,50
79	não metrop.	4,30	855,0	285,60	356,80	44,59	1700	42,00
80 = hh	metrop.	4,40	850,0	256,10	348,50	42,20	227	39,00
81 = e	não metrop.	4,00	859,0	290,20	362,40	44,30	1700	50,50
82*	não metrop.	2,60	854,0	283,00	353,10	44,50	1700	39,00
83*	não metrop.	3,30	854,0	285,00	357,30	44,70	1700	43,50
84	não metrop.	3,70	854,0	286,40	357,40	44,80	1600	42,00
85 = k*	não metrop.	3,00	853,0	278,70	345,60	44,50	1400	45,00
86	não metrop.	3,50	854,0	287,40	358,00	45,50	1600	45,00
87 = f*	metrop.	8,40	850,0	268,20	349,50	43,90	300	44,00
88	não metrop.	4,30	852,0	280,20	354,50	44,70	1500	41,00
89 = mm*	não metrop.	3,20	852,0	284,80	359,80	45,40	1600	42,00
90 = v*	não metrop.	3,30	854,0	285,10	355,70	44,70	1600	40,50
91 = g*	não metrop.	10,20	857,0	291,50	349,90	45,10	1300	45,50
92*	não metrop.	3,20	853,0	281,70	356,80	44,60	1600	41,00
93*	não metrop.	3,20	854,0	280,90	351,80	44,20	1600	42,00
94 = h*	metrop.	8,30	847,0	265,70	350,80	44,89	100	41,00
95 = nn*	não metrop.	3,40	855,0	286,30	354,60	45,00	1400	44,00
96	metrop.	4,40	848,0	253,80	347,80	42,20	100	38,50
97 = i*	metrop.	7,20	848,0	264,50	347,60	44,10	200	41,00
98 = ee	metrop.	4,20	849,0	261,00	355,40	43,30	300	39,00
99	não metrop.	3,50	854,0	280,20	352,10	44,30	1700	40,50
100 = aa*	não metrop.	3,20	854,0	286,80	358,20	45,00	1700	43,00

[a] Teor de biodiesel como % em volume (média = 3,85%, desvio padrão = 1,11%), determinado pela norma ABNT NBR 15568:2008 (Biodiesel - Determinação do teor de biodiesel em óleo diesel por espectroscopia na região do infravermelho médio), padrão (na época em que as amostras foram coletadas): 4,0 ± 0,5 %; [b] Massa específica a 20 °C (média = 852,1 Kg/m^3, desvio padrão = 3,13 Kg/m^3), padrão: amostras metropolitanas entre 820 e 865 Kg/m^3; amostras não metropolitanas entre 820 e 880 Kg/m^3; [c] Temperatura em que 50% do volume é destilado (média = 272,12 °C, desvio padrão = 14,06 °C), padrão: entre 245 e 310 °C; [d] Temperatura em que 85% do volume é destilado (média = 352,46 °C, desvio padrão = 5,10 °C), padrão: amostras metropolitanas, máximo de 360 °C; amostras não metropolitanas, máximo de 370 °C; [e] Índice de cetano (adimensional, média =43,73, desvio padrão = 1,23), padrão: mínimo de 42; [f] Teor de enxofre em mg/Kg (média = 886 mg/Kg, desvio padrão = 681 mg/Kg), padrão (na época em que as amostras foram coletadas): amostras metropolitanas, máximo de 500 mg/Kg; amostras não metropolitanas, máximo de 1800 mg/Kg; [g] Ponto de fulgor em °C (média = 42,03 °C, desvio padrão = 3,36 °C), padrão: mínimo de 38 °C. Os valores em vermelho encontram-se fora da especificação.

Tabela 4: Numeração dos *buckets* e suas faixas em ppm nos espectros de RMN ^{1}H.

Nº	Ppm	Nº	ppm	Nº	ppm	Nº	ppm	Nº	ppm
1	10,0-9,97	39	7,94-7,89	77	5,99-5,94	115	4,06-3,99	153	2,01-2,00
2	9,97-9,90	40	7,89-7,84	78	5,94-5,90	116	3,99-3,96	154	2,00-1,91
3	9,90-9,86	41	7,84-7,79	79	5,90-5,79	117	3,96-3,85	155	1,91-1,88
4	9,86-9,77	42	7,79-7,74	80	5,79-5,77	118	3,85-3,84	156	1,88-1,87
5	9,77-9,72	43	7,74-7,69	81	5,77-5,75	119	3,84-3,81	157	1,87-1,82
6	9,72-9,66	44	7,69-7,64	82	5,75-5,70	120	3,81-3,76	158	1,82-1,77
7	9,66-9,61	45	7,64-7,59	83	5,70-5,59	121	3,76-3,71	159	1,77-1,72
8	9,61-9,53	46	7,59-7,54	84	5,59-5,59	122	3,71-3,66	160	1,72-1,61
9	9,53-9,46	47	7,54-7,47	85	5,59-5,55	123	3,66-3,61	161	1,61-1,58
10	9,46-9,38	48	7,47-7,44	86	5,55-5,50	124	3,61-3,56	162	1,58-1,57
11	9,38-9,36	49	7,44-7,33	87	5,50-5,39	125	3,56-3,50	163	1,57-1,52
12	9,36-9,29	50	7,33-7,30	88	5,39-5,35	126	3,50-3,46	164	1,52-1,42
13	9,29-9,25	51	7,30-7,29	89	5,35-5,30	127	3,46-3,35	165	1,42-1,37
14	9,25-9,19	52	7,29-7,24	90	5,30-5,19	128	3,35-3,34	166	1,37-1,32
15	9,19-9,16	53	7,24-7,13	91	5,19-5,14	129	3,34-3,31	167	1,32-1,27
16	9,16-9,13	54	7,13-7,13	92	5,14-5,12	130	3,31-3,26	168	1,27-1,16
17	9,13-9,03	55	7,13-7,09	93	5,12-5,10	131	3,26-3,21	169	1,16-1,11
18	9,03-8,98	56	7,09-7,02	94	5,10-4,99	132	3,21-3,16	170	1,11-1,06
19	8,98-8,91	57	7,02-6,99	95	4,99-4,97	133	3,16-3,11	171	1,06-1,01
20	8,91-8,84	58	6,99-6,94	96	4,97-4,95	134	3,11-3,06	172	1,01-0,98
21	8,84-8,81	59	6,94-6,89	97	4,95-4,90	135	3,06-2,98	173	0,98-0,97
22	8,81-8,73	60	6,89-6,84	98	4,90-4,80	136	2,98-2,96	174	0,97-0,92
23	8,73-8,69	61	6,84-6,74	99	4,80-4,76	137	2,96-2,91	175	0,92-0,87
24	8,69-8,63	62	6,74-6,69	100	4,76-4,73	138	2,91-2,82	176	0,87-0,76
25	8,63-8,58	63	6,69-6,64	101	4,73-4,70	139	2,82-2,81	177	0,76-0,71
26	8,58-8,54	64	6,64-6,59	102	4,70-4,66	140	2,81-2,76	178	0,71-0,66
27	8,54-8,49	65	6,59-6,54	103	4,66-4,55	141	2,76-2,67	179	0,66-0,61
28	8,49-8,44	66	6,54-6,49	104	4,55-4,54	142	2,67-2,66	180	0,61-0,56
29	8,44-8,37	67	6,49-6,44	105	4,54-4,48	143	2,66-2,56	181	0,56-0,51
30	8,37-8,34	68	6,44-6,39	106	4,48-4,45	144	2,56-2,54	182	0,51-0,46
31	8,34-8,23	69	6,39-6,34	107	4,45-4,41	145	2,54-2,51	183	0,46-0,42
32	8,23-8,18	70	6,34-6,29	108	4,41-4,35	146	2,51-2,41	184	0,42-0,37
33	8,18-8,13	71	6,29-6,19	109	4,35-4,30	147	2,41-2,36	185	0,37-0,32
34	8,13-8,10	72	6,19-6,14	110	4,30-4,26	148	2,36-2,26	186	0,32-0,27
35	8,10-8,09	73	6,14-6,09	111	4,26-4,20	149	2,26-2,22	187	0,27-0,22
36	8,09-8,04	74	6,09-6,08	112	4,20-4,16	150	2,22-2,12	188	0,22-0,17
37	8,04-7,98	75	6,08-6,05	113	4,16-4,09	151	2,12-2,06	189	0,17-0,12
38	7,98-7,94	76	6,05-5,99	114	4,09-4,06	152	2,06-2,01	190	0,12-0,07

Para a aplicação da DWT, as diferentes funções de uma família *wavelet* são especificadas por vetores contendo um conjunto de números denominados coeficientes filtros *wavelet*, de onde são definidos dois filtros digitais, a saber, o filtro passa-baixa (*low-pass filter*), definido por uma função acessório (também derivada da função mãe), chamada função de escalamento (ϕ), e o filtro passa-alta (*high-pass filter*). Os filtros são usados para construir as matrizes filtro, denotadas por **G** e **H**, respectivamente. Ambos os filtros usam o mesmo conjunto de coeficientes, mas em ordem inversa e com sinais alternados. O filtro **G** fornece a parte do sinal com menor frequência, calculando os chamados coeficientes de aproximação. Do ponto de vista químico, essa parte do sinal (espectro) corresponde às características médias, como linha de base, por exemplo. O filtro **H** fornece a parte do sinal com maior frequência, denominados coeficientes de detalhes e que quimicamente representam as características bem localizadas do sinal, tais como picos estreitos. Portanto, a decomposição por DWT é capaz de extrair características localizadas do sinal no plano tempo-frequência, concentrando as informações determinísticas em poucos coeficientes (graças à compactação de energia das *wavelets*) e de certa forma espalhando as características estocásticas (como o ruído aleatório) através de todos os outros coeficientes. Deste modo, as operações de compressão e de diminuição de ruído podem ser realizadas pela reconstrução do sinal após a eliminação dos coeficientes que não retém as informações importantes. Na reconstrução do sinal, os filtros inversos são utilizados, compondo as inversas das matrizes **G** e **H** que por serem ortogonais tem inversa igual à matriz transposta [95].

A Figura 16 mostra um exemplo de função *wavelet* (da família Daubechies, cujo símbolo é db10) e a função de escalamento associada, bem como os filtros de decomposição e de reconstrução derivados destas funções. Nesta figura, nota-se a já citada assimetria das funções e a concentração de energia das mesmas (os maiores valores estão numa região limitada). Os filtros são colocados em diferentes posições

nas linhas das matrizes **G** e **H,** que possuem número de linhas apropriado para cobrir todo o sinal analisado. As demais posições em cada linha são preenchidas com zeros. Portanto, as dimensões das matrizes dependem da dimensão do sinal analisado. Vale ressaltar que estas matrizes são também chamadas de filtros espelhos em quadratura (QMF, do inglês *Quadrature Mirror Filters*).

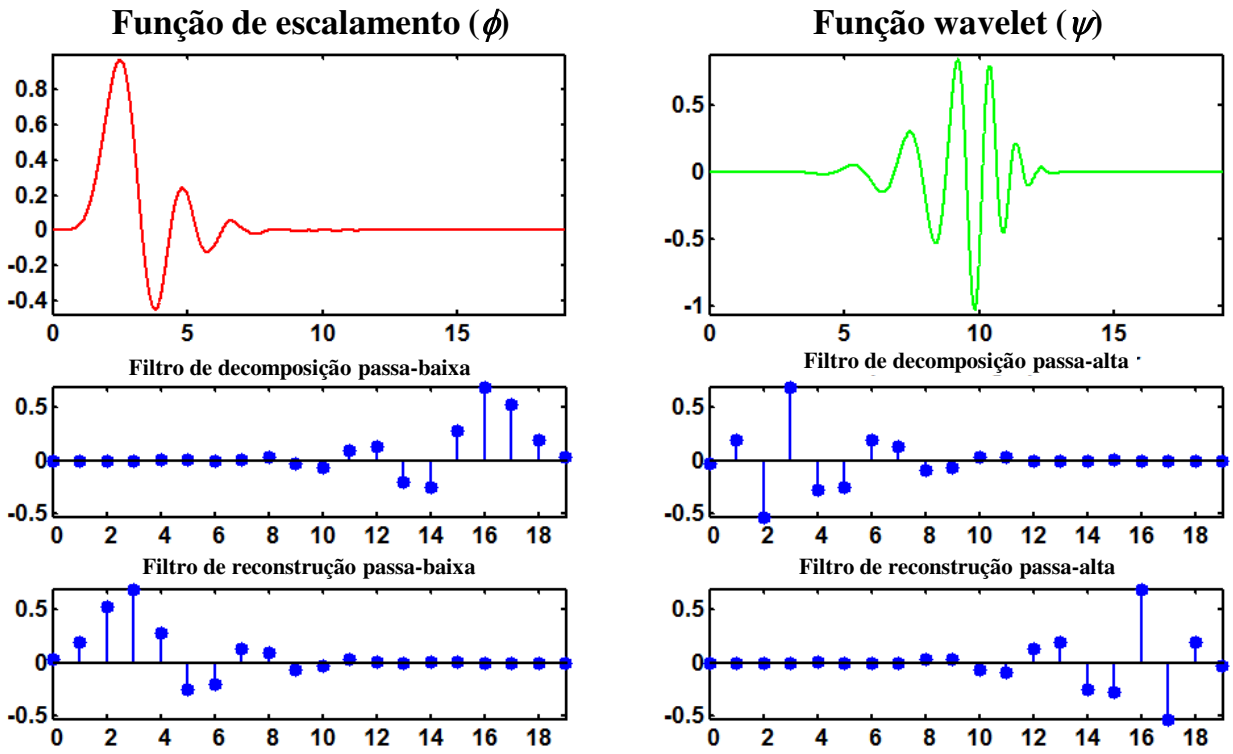

Figura 16: Exemplo de função *wavelet* da família Daubechies. A função é simbolizada por db10. Em todos os gráficos têm-se nas abcissas os índices dos coeficientes (que neste caso vai de 0 até 19, ou seja, 20 coeficientes) e nas ordenadas os seus valores.

Na aplicação do QMF num sinal com N elementos de entrada (por exemplo, um espectro de RMN com 32768 pontos), ou seja, um vetor $\mathbf{f}$ ($N \times 1$) gera-se uma saída também com N valores, porém $N/2$ em coeficientes de aproximação, ou seja, um vetor $\mathbf{a}^1$ ($N/2 \times 1$), e $N/2$ em coeficientes de detalhes, ou seja, um vetor $\mathbf{d}^1$ ($N/2 \times 1$). Os coeficientes de aproximação podem ser utilizados como entradas para um novo QMF

gerando $N/4$ novos coeficientes de aproximação, num vetor $\mathbf{a}^2$ ($N/4 \times 1$) e $N/4$ novos coeficientes de detalhes, num vetor $\mathbf{d}^2$ ($N/4 \times 1$), em um segundo nível de resolução. Portanto, a cada filtragem o sinal sofre uma diminuição em metade dos pontos (os elementos com índices pares, 0, 2, 4,... são removidos), sendo esta operação conhecida como decimação ou subamostragem [90,93,94]. A cada nível a aproximação fica mais suavizada, dada a remoção contínua de detalhes. O processo pode prosseguir até alcançar um intervalo unitário, porém pode ser interrompido a qualquer nível de resolução, seguindo algum critério pré-estabelecido. Todos os coeficientes obtidos são indexados pelos parâmetros diáticos.

As fórmulas recursivas na Equação 18 resumem o algoritmo DWT proposto por Mallat [90] e também conhecido como algoritmo pirâmide:

$$\mathbf{a}^{m+1} = \mathbf{G}\mathbf{f}^m \qquad \mathbf{d}^{m+1} = \mathbf{H}\mathbf{f}^m \qquad \text{Equação 18}$$

Onde m denota o nível de resolução. Quando $m = 0$, o vetor $\mathbf{f}$ ($N \times 1$) representa o sinal original e quando $m = 1, 2, 3, ..., M$, o vetor $\mathbf{f}$ representa os coeficientes de aproximação de um nível anterior, sendo M o número total de níveis. As matrizes $\mathbf{G}$ e $\mathbf{H}$ mudam a cada nível de decomposição, pois suas dimensões adaptam-se ao sinal analisado e seguem a relação $\mathbf{G}_m$ ($N/2^m \times N$) e $\mathbf{H}_m$ ($N/2^m \times N$). Os vetores $\mathbf{a}^{m+1}$ ($N/2^{(m+1)} \times 1$) e $\mathbf{d}^{m+1}$ ($N/2^{(m+1)} \times 1$) denotam os coeficientes de aproximação e detalhes, respectivamente, e onde as dimensões já consideram a etapa de decimação. A Figura 17 mostra um esquema da execução do algoritmo pirâmide, onde podem ser notadas as operações de decimação a cada etapa [90].

O número de níveis e a melhores famílias *wavelet* (funções de base) utilizados na decomposição por DWT de algum sinal podem ser determinados pelo critério de entropia. As bases mais eficientes devem fornecer na decomposição alguns coeficientes retendo altos valores (informação relevante) e os remanescentes retendo

baixos valores, ou seja, a maior diferenciação possível dentro do conjunto de coeficientes. Quando os coeficientes têm mais ou menos os mesmos valores, há baixa concentração de informação relevante e em outras palavras alta entropia. Portanto, a base que fornece o máximo de informação é referente ao mínimo de entropia para a distribuição dos coeficientes [94].

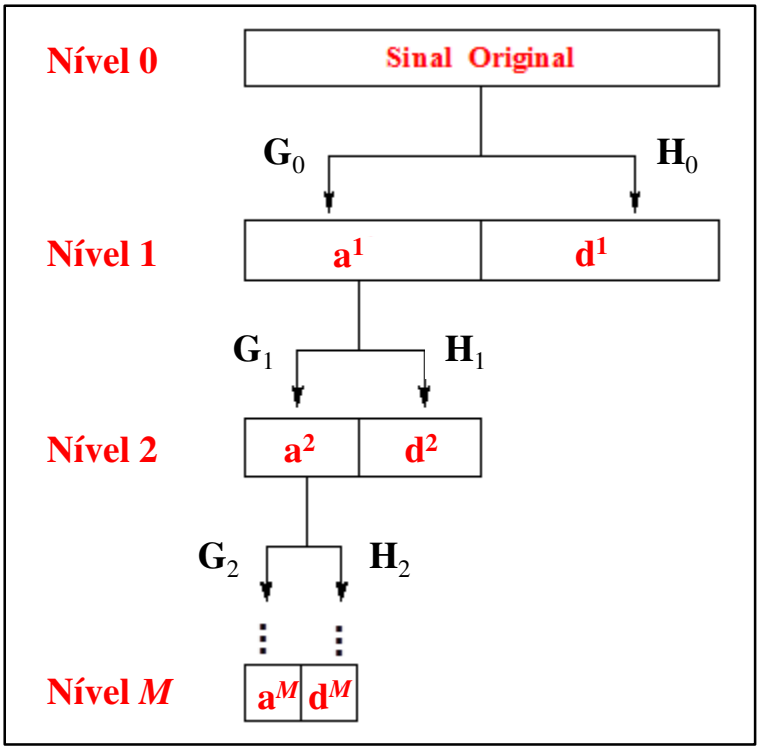

Figura 17: Representação esquemática do algoritmo pirâmide para execução da DWT.

Neste trabalho, utilizou-se a entropia de Shannon como critério de entropia e as famílias de *wavelets* avaliadas foram Daubechies e Symlets, sendo dez bases ortonormais (db1 – 10) da primeira e sete bases ortonormais (sym1 – 7) da segunda. Os números que acompanham os símbolos (db e sym) das *wavelets* indicam as ordens das funções, ou seja, a metade do número de elementos de cada vetor filtro [94].

Na Equação 19 o valor de s_i corresponde à variância de cada variável da matriz de sinais multivariados (espectros de RMN ^{13}C), aqui estudada, sendo esta

representada pelo vetor de variâncias tanto nas decomposições quanto no cálculo da entropia de Shannon [94].

$$Entr.Shan. = -\sum_{i=1}^{N} s_i^2 \log(s_i^2) \qquad \text{Equação 19}$$

Como pode ser visto na Figura 18, a decomposição dos sinais até o segundo nível leva a uma diminuição da entropia de Shannon e a partir deste nível a entropia aumenta para todas as bases, não sendo assim interessante a decomposição no terceiro nível. Portanto, as bases com menor entropia foram a db7 e a sym8 a dois níveis de decomposição, sendo assim arbitrariamente escolhida a função de base db7 para os estudos posteriores deste trabalho.

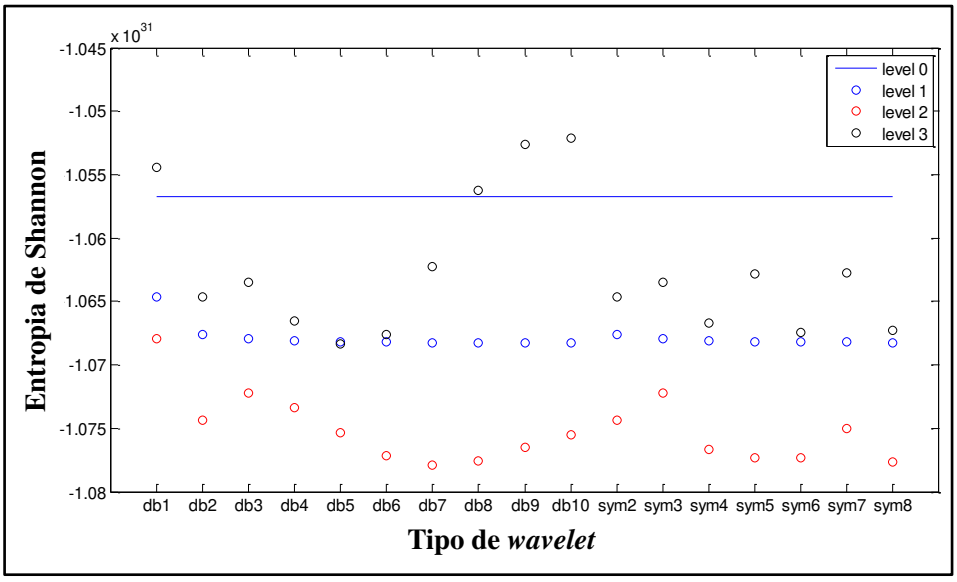

Figura 18: Valores de entropia de Shannon calculados para a matriz original (— nível zero) e para as decomposições nos níveis um (o), dois (o) e três (o) usando os componentes das famílias Daubechies e Symlets. A matriz com os espectros de RMN ^{13}C é representada pelo vetor de variâncias de cada variável.

3.4.4.2. Análise de componentes principais em múltiplas escalas

Iniciando com uma matriz de sinais multivariados (neste caso, a matriz com os espectros de RMN ^{13}C), a análise de componentes principais em múltiplas escalas (MSPCA) reconstrói os sinais multivariados simplificados utilizando uma PCA sem centragem na média sobre os coeficientes de aproximações e de detalhes obtidos numa decomposição por DWT e uma PCA final. A cada nível de decomposição, as componentes principais mais significantes são retidas e assim sinais simplificados interessantes podem ser reconstruídos com apreciável efeito de diminuição de ruído. O MSPCA combina a habilidade da PCA em descorrelacionar as variáveis através da extração de relações lineares com a habilidade da TW de extrair características determinísticas usando uma representação simples dos dados em cada nível de resolução [94,96].

Neste trabalho, uma vez determinado o número de níveis de decomposição como citado acima (dois níveis de decomposição), a MSPCA foi aplicada ao conjunto de dados, numa matriz $\mathbf{X}$ ($I \times J$) composta pelos espectros de RMN ^{13}C, seguindo o esquema que é mostrado na Figura 19, onde podem ser vistas todas as etapas a que cada matriz é submetida, o número de componentes utilizados nas reconstruções após PCA, bem como as dimensões das matrizes envolvidas. No final, uma matriz simplificada com diminuição de ruído e com as mesmas dimensões de $\mathbf{X}$, é obtida.

Na MSPCA, o número de componentes principais para a compressão deve ser cuidadosamente escolhido, buscando um equilíbrio entre a supressão de ruído e o sinal remanescente. A escolha de um número baixo de componentes principais pode levar à perda de sinais importantes. Neste trabalho, para os detalhes no primeiro e no segundo nível de decomposição, cinco e quatro componentes principais foram usadas, respectivamente, sendo estes números escolhidos baseando-se na regra de Kaiser, onde os componentes associados com os autovalores que excedem a média de

autovalores são mantidos. Na PCA final e na aproximação no segundo nível de decomposição, seis componentes principais foram utilizadas [94,96].

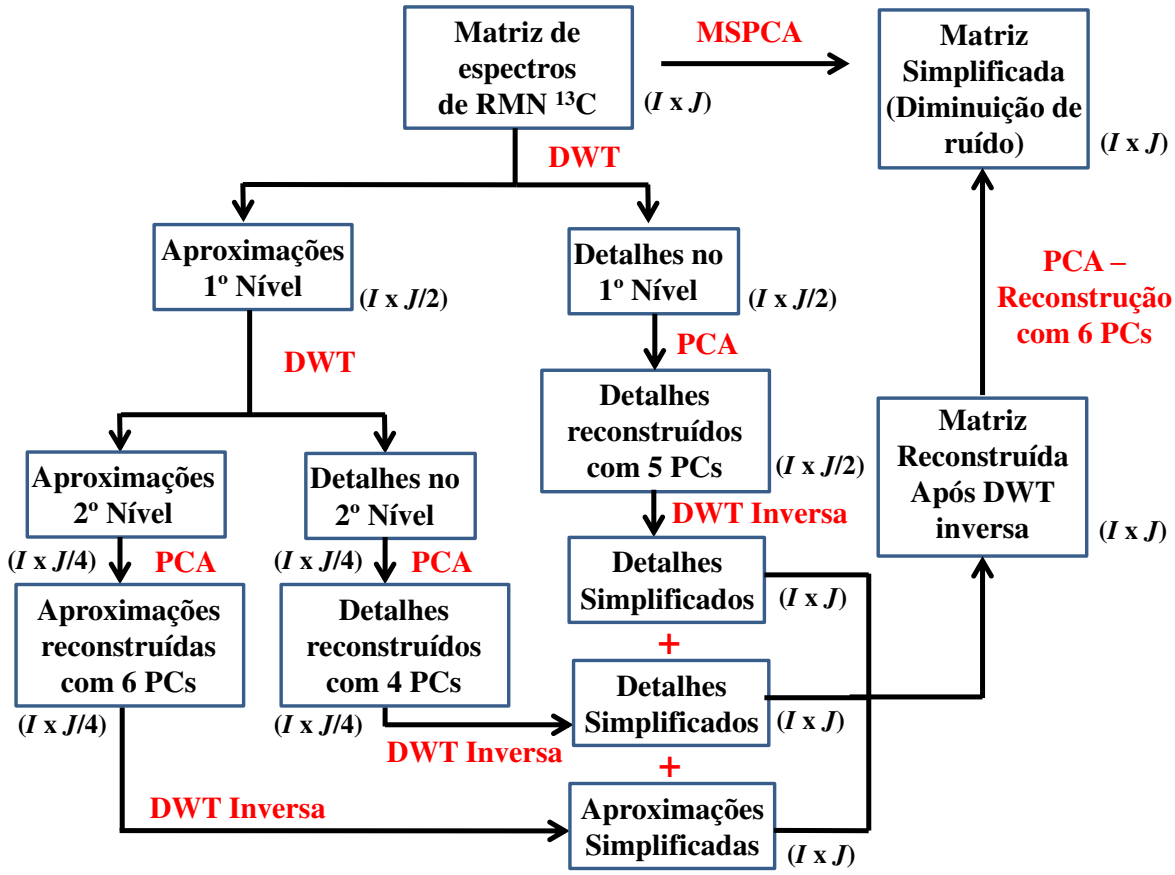

Figura 19: Esquema para execução da MSPCA neste trabalho. As dimensões de todas as matrizes envolvidas são indicadas.

3.4.5. Análise quimiométrica

A matriz **Z** (100 × 191) obtida após o pré-tratamento com o método de *bucketing* otimizado (OBA) dos espectros de RMN ^{1}H e a matriz de espectros de RMN ^{13}C após a diminuição de ruído e de dimensão **X** (100 × 22680) relativa à região entre 0,1 e 180 ppm, foram centradas na média e submetidas a análises de

componentes principais em intervalos (iPCA) usando rotinas da iToolbox para Matlab. Os intervalos estudados são mostrados na Tabela 5 e foram determinados baseando-se nas atribuições dos sinais dos espectros de RMN, sendo evitadas regiões com apenas linha de base (note que, por este motivo, os intervalos na Tabela 5 podem não cobrir a faixa inteira). Elipsoides com 95% de intervalo de confiança baseados na estatística T^2 de Hotelling foram construídos para cada agrupamento através de uma rotina escrita em código Matlab (The MathWorks, Natick, MA, USA). A base matemática para a construção dos elipsoides é descrita a seguir.

Tabela 5: Intervalos dos espectros usados para construção dos modelos de iPCA.

Nº do Intervalo	Região do espectro de ^{1}H		Região do espectro de ^{13}C (ppm)
	ppm	Intervalos dos *buckets*	
1	10,0 – 8,1	1 – 35	173,0 – 171,3
2	8,1 – 7,1	36 – 54	145,1 – 132,4
3	7,1 – 6,1	55 – 75	132,4 – 116,6
4	5,6 – 5,0	84 – 96	51,2 – 49,1
5	4,2 – 3,7	112 – 122	49,1 – 40,0
6	3,7 – 3,2	123 – 131	40,0 – 32,5
7	3,2 – 2,8	132 – 139	32,5 – 30,4
8	2,8 – 2,5	140 – 144	30,4 – 28,5
9	2,5 – 2,2	145 – 149	28,5 – 26,0
10	2,2 – 2,0	150 – 153	26,0 – 24,4
11	2,0 – 1,8	154 – 157	24,4 – 21,8
12	1,8 – 1,4	158 – 165	21,8 – 17,0
13	1,4 – 1,0	166 – 173	17,0 – 15,0
14	1,0 – 0,3	174 – 185	15,0 – 11,5
15	-	-	11,5 – 8,0

A matriz de dados referente aos parâmetros de qualidade das amostras (Tabela 3) foi autoescalada e submetida à análise por componentes principais (PCA) utilizando o *software* Pirouette 3.11 (Infometrix, Seattle, WA, USA). Do mesmo

modo, as matrizes completas (sem o uso de intervalos) dos espectros de RMN ^{1}H e ^{13}C foram submetidas à análise de componentes principais (PCA) com os dados centrados na média.

3.4.5.1. Cálculo dos elipsoides com 95% de confiança

A estatística T^2 de Hotelling pode ser considerada a extensão multivariada da estatística *t* univariada de Student [97]. Dada uma matriz de escores não normalizados **T** ($I \times R$), onde R componentes principais são utilizadas, sendo a matriz **T** obtida na PCA de uma matriz **X** ($I \times J$) com I amostras independentemente amostradas e J variáveis normalmente distribuídas, calcula-se a matriz de covariância **S** ($R \times R$) pela Equação 20:

$$\mathbf{S} = \frac{1}{I-1}(\mathbf{T} - \mathbf{1}\bar{\mathbf{t}}^{\mathrm{T}})^{\mathrm{T}}(\mathbf{T} - \mathbf{1}\bar{\mathbf{t}}^{\mathrm{T}}) \qquad \text{Equação 20}$$

Onde **1** é um vetor coluna de uns com dimensão ($I \times 1$) e $\bar{\mathbf{t}}$ é um vetor coluna com dimensão ($R \times 1$) contendo as médias de cada uma das colunas da matriz **T**.

Sejam agora os escores de uma amostra futura, contidos num vetor **z** ($R \times 1$), obtidos pela projeção desta amostra no subespaço gerado pelas R componentes principais decorrentes da PCA da matriz **X** ($I \times J$), sendo medidas sobre a amostra futura, as mesmas J variáveis de **X**. O desvio dos escores desta amostra em relação à média dos escores em $\bar{\mathbf{t}}$ pode ser medido através do parâmetro T^2 de Hotelling para esta amostra que é calculado como mostrado na Equação 21:

$$T^2 = \left(\frac{I}{I+1}\right)(\mathbf{z}-\bar{\mathbf{t}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{z}-\bar{\mathbf{t}}) \qquad \text{Equação 21}$$

Onde **S** é a matriz de covariância calculada através das amostras passadas (amostras em **X**) como mostrado na Equação 20. A razão $I/(I + 1)$ corresponde a um termo de normalização da matriz de covariância **S**, dada a inclusão da amostra futura [98,99].

O parâmetro T^2 de Hotelling tem distribuição F com R graus de liberdade no numerador e $(I - R)$ graus de liberdade no denominador. Desse modo, para a amostra futura pertencer à mesma população das amostras na matriz **X** é necessário que:

$$\left(\frac{I}{I+1}\right)(\mathbf{z}-\bar{\mathbf{t}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{z}-\bar{\mathbf{t}}) \leq \frac{R(I-1)}{I-R}F_{(\alpha,R,I-R)} \qquad \text{Equação 22}$$

Onde $F_{(\alpha,R,I-R)}$ corresponde ao valor de F crítico na distribuição F para o nível de significância α e R graus de liberdade no numerador e $(I - R)$ graus de liberdade no denominador. O termo $(I - 1)$ corresponde ao número de graus de liberdade no cálculo de **S** e é incluído tendo em vista que esta matriz é uma estimativa da matriz de covariância verdadeira para a população de amostras e deste modo deve ser usada como um estimador não tendencioso (*unbiased*) [98,99]. Ainda em relação à Equação 22, em outras palavras o que está sendo testado é uma hipótese nula (H_0) de que não há diferença entre os escores da amostra futura e a média dos escores obtidos das amostras passadas contra uma hipótese alternativa (H_A) de que os escores da amostra futura são significativamente diferentes da média de escores das amostras passadas e assim esta amostra futura não pertence à mesma população das amostras. Adicionalmente, a questão pode ser vista como uma H_0 de que a diferença entre os escores da amostra futura e a média dos escores das amostras passadas segue uma distribuição multinormal (pressuposta na dedução da estatística T^2 de Hotelling) *R*-

dimensional com média zero e matriz de covariância proporcional à matriz de covariância verdadeira [98,99]. Desse modo é necessário rearranjar a Equação 22 para que o teste seja feito no espaço dos escores. Assim, obtemos:

$$(\mathbf{z}-\bar{\mathbf{t}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{z}-\bar{\mathbf{t}}) \leq \frac{R(I-1)(I+1)}{I(I-R)} F_{(\alpha,R,I-R)} \qquad \text{Equação 23}$$

Para evidenciar, a partir da Equação 23, a distância entre os escores da amostra futura e o centroide dos escores das amostras passadas devemos fazer uso da decomposição de valores singulares (SVD) da matriz de covariância **S**. Portanto, temos:

$$\mathbf{S} = \mathbf{U}\mathbf{D}\mathbf{P}^{\mathrm{T}} \qquad \text{Equação 24}$$

Vale ressaltar que **S** é uma matriz simétrica definida positiva e tem inversa, uma vez que os escores não normalizados em **T** são não correlacionados, assim **U** ($R \times R$) = **P** ($R \times R$), ou seja, o espaço coluna da matriz **S** é igual ao espaço linha, sendo ambas as matrizes ortonormais, valendo as relações $\mathbf{U}^{-1} = \mathbf{U}^{\mathrm{T}}$ e $\mathbf{P}^{-1} = \mathbf{P}^{\mathrm{T}}$, além de $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{P}\mathbf{P}^{\mathrm{T}} = \mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I}$ ($R \times R$). **D** é uma matriz diagonal com os autovalores de **S** todos positivos e correspondentes à variância dos dados em cada uma das componentes principais. Nota-se que os autovalores obtidos na SVD de **S** fornecem diretamente a variância na matriz $\widehat{\mathbf{X}}$ reconstruída pelas R componentes principais obtidas na PCA, cujos escores não normalizados das I amostras estão presentes em **T**, uma vez que **S** é a matriz de covariância associada a essa matriz reconstruída. Comumente, na PCA através de SVD, é necessário tomar o quadrado dos autovalores para calcular a variância, uma vez que a SVD é computada diretamente da matriz de dados e não da matriz de covariância. Diante disto temos:

$$\mathbf{S}^{-1} = (\mathbf{U}^{\mathrm{T}})^{-1}\mathbf{D}^{-1}(\mathbf{U})^{-1} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}^{\mathrm{T}} = \sum_{r=1}^{R} \frac{\mathbf{u}_r \mathbf{u}_r^{\mathrm{T}}}{\lambda_r} \qquad \text{Equação 25}$$

Onde $r = 1, ..., R$ e:

$$\mathbf{D}^{-1} = \begin{bmatrix} 1/\lambda_1 & & & \\ & 1/\lambda_2 & & \\ & & \ddots & \\ & & & 1/\lambda_R \end{bmatrix} \qquad \text{Equação 26}$$

Onde os termos λ_i correspondem aos autovalores de **S**.

Considerando o teste de hipótese num espaço de escores com duas componentes principais ($R = 2$), um nível de significância $\alpha = 0{,}05$, substituindo as Equações 25 e 26 na Equação 23 e fazendo uso do índice r, onde $r = 1, ..., R$, obtém-se a Para $R = 2$. Nesta equação o vetor $\mathbf{u}_r$ tem dimensão ($R \times 1$) e corresponde ao autovetor associado à r-ésima componente principal. Este vetor é normalizado e, portanto, a operação do lado esquerdo da Para $R = 2$ fornece o quadrado da diferença entre os escores da amostra futura e o centroide dos escores das amostras passadas. No caso da matriz inicial **X** ($I \times J$) ser centrada na média, os escores das I amostras também serão, logo o lado esquerdo da Para $R = 2$ fornecerá o quadrado dos escores da amostra futura diretamente.

$$\sum_{r=1}^{R} (\mathbf{z} - \bar{\mathbf{t}})^{\mathrm{T}} \mathbf{u}_r \mathbf{u}_r^{\mathrm{T}} (\mathbf{z} - \bar{\mathbf{t}}) \leq \frac{2\lambda_r (I^2 - 1)}{I(I-2)} F_{(0{,}05,2,I-2)} \quad \text{Para } R = 2 \qquad \text{Equação 27}$$

A hipótese nula H_0 é rejeitada se o parâmetro calculado do lado esquerdo da Para $R = 2$ for maior do que o valor calculado do lado direito da mesma equação para

um valor de F crítico a uma significância de $\alpha = 0{,}05$ e graus de liberdade 2 e $(I - 2)$ no numerador e no denominador, respectivamente. Portanto, a raiz quadrada do valor do lado direito da equação fornece o valor máximo de escore para que uma amostra faça parte da mesma distribuição multinormal das demais, num intervalo de confiança de 95% para a r-ésima componente principal. Esse valor define a metade do eixo de um elipsoide ao longo da componente principal r [98,99].

Neste trabalho, elipsoides com 95% de confiança são traçados em gráficos bidimensionais com base na teoria acima, servindo para a detecção de amostras fora da especificação, ou seja, aquelas amostras situadas fora do elipsoide e que assim diferem significativamente das demais. Adicionalmente, quando as variabilidades das duas diferentes classes de amostras, aqui estudadas, são incluídas num mesmo modelo PCA, são calculados dois elipsoides, sendo uma para cada classe. Os elipsoides são centrados na média da sua classe e inclinados de acordo com o coeficiente angular obtido na regressão dos escores de uma componente sobre a outra, que quantifica a covariância entre os subgrupos de escores.

3.5. Resultados e Discussão

3.5.1. Atribuições dos espectros de RMN de ^{1}H e ^{13}C

Os espectros de RMN ^{1}H, numa matriz **X** (100 × 15850), e de RMN ^{13}C, numa matriz **X** (100 × 22680), das cem misturas de biodiesel-diesel são mostrados nas Figuras 20A e 20B, onde os espectros de ^{1}H entre 0,1 e 2,3 ppm e os espectros de ^{13}C entre 5,0 e 45 ppm podem ser vistos como regiões ampliadas, sendo mostrados sem nenhum pré-tratamento. Os desalinhamentos percebidos na região ampliada dos espectros de RMN ^{1}H na Figura 20A foram corrigidos segundo descrito no capítulo 2, fornecendo uma matriz de *buckets* **Z** (100 × 191). Os espectros de RMN ^{13}C foram submetidos à redução de ruído utilizando o método MSPCA, conforme descrito nos itens 3.4.4.1 e 3.4.4.2. A matriz resultante também tem dimensão (100 × 22680).

Nos espectros de RMN ^{1}H, entre os picos relativos ao biodiesel podem ser destacados os singletos entre 3,4 e 3,5 ppm, que correspondem aos hidrogênios metílicos nas porções éster e os singletos entre 5,3 e 5,4 ppm que referem-se aos hidrogênios dos grupos olefínicos nas cadeias longas de alguns dos ésteres. Os outros deslocamentos químicos do biodiesel estão sobrepostos com os picos relacionados ao diesel. Estes sinais são dos hidrogênios alifáticos na cadeia dos ésteres e ocorrem entre 0,7 e 3,0 ppm, por exemplo, em 2,4 e 1,6 ppm os hidrogênios nas posições α e β em relação ao grupo carbonila do éster, respectivamente, e em 2,0 e 2,7 ppm, os hidrogênios nas posições alílica e *bis*-alílica, respectivamente. Na faixa de 6,4-9,0 ppm há vários picos de hidrogênios aromáticos em compostos mono e polinucleares (substituídos ou não-substituídos) que constituem o óleo diesel e talvez os corantes, utilizados para a diferenciação por cor entre o diesel comercializado nas regiões metropolitana e não-metropolitana. Os picos de 2,0 a 3,0 ppm são atribuídos a grupos diretamente ligados a anéis aromáticos, enquanto os dois picos a esquerda dos sinais

mais intensos, entre 1,6 e 1,9 ppm e 1,4 e 1,6 ppm referem-se aos naftenos e *iso*-parafinas, respectivamente. Os dois maiores picos de 1,1 a 1,3 ppm e de 0,7 a 0,9 ppm correspondem a grupos CH_2 e CH_3, respectivamente, de normal (*n*-) ou *iso*-parafinas [82,100,101].

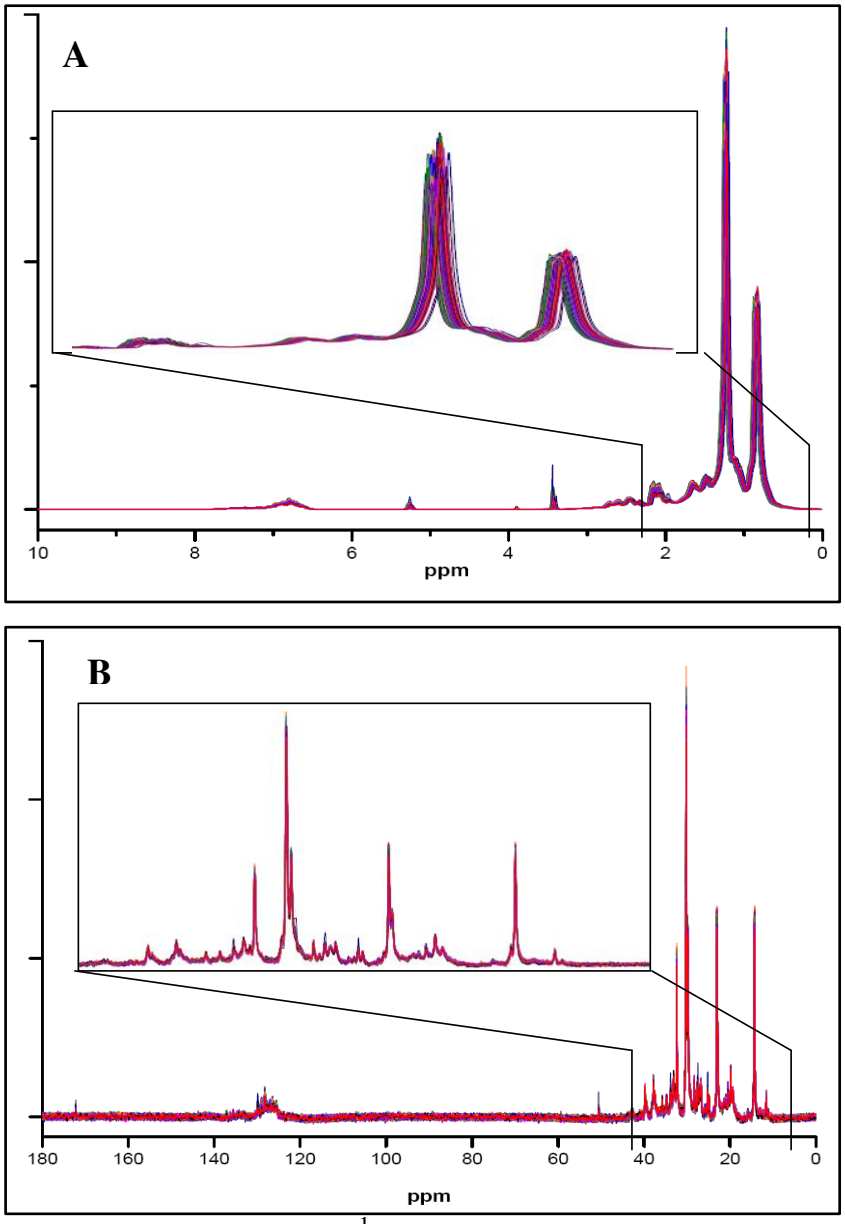

Figura 20: A: Espectros de RMN ^{1}H das misturas de biodiesel-diesel. Região ampliada entre 0,1 e 2,3 ppm. B: Espectros de RMN ^{13}C das misturas de biodiesel-diesel. Região ampliada entre 5,0 e 45 ppm.

Nos espectros de RMN ^{13}C, os maiores picos ocorrem em 23,0; 29,8; 30,2 e 32,3 ppm e estão relacionados com os grupos metilenos das cadeias das *n*-parafinas, enquanto que o pico a 14,2 ppm corresponde ao CH_3, no final das mesmas cadeias. Outros sinais menos intensos são ligados a grupos metilenos nas *iso*-parafinas e naftenos, tais como os picos em 19,8; 27,5 e 37,5 ppm. Além disso, existem alguns picos entre 10,0 e 22,0 ppm atribuídos aos átomos de carbono das extremidade de várias ramificações nas *iso*-parafinas, em 11,0; 11,5; 14,6 e 19,4 ppm. Estas ramificações podem ser metilas, etilas, propilas, butilas, entre outras. Os grupos metilenos destas ramificações podem ser observados em 20,4; 21,0 e 21,4 ppm. Os deslocamentos químicos em 24,8; 25,3; 26,7; 27,2; 27,6; 28,3; 33,1 e 33,8 ppm são atribuídos à cadeia longa dos ésteres que constituem o biodiesel, onde também surgem os picos a 50,5 ppm (metilas nas porções éster); 129,9 e 128,2 ppm (carbonos olefínicos da cadeia dos ésteres) e em 172,2 ppm (carbonilas dos ésteres). Os sinais em 32,7; 34,7 e 35,7 estão relacionados com os átomos de carbono ligados a carbonos olefínicos e aqueles em 37,8; 39,5; e 39,8 ppm referem-se a átomos de carbono ligados a núcleos aromáticos. Na faixa de 110 a 145 ppm, existem cerca de dez picos de átomos de carbono em núcleos aromáticos [102,103].

3.5.2. Análise exploratória por iPCA: espectros de RMN ^{1}H

No presente trabalho, abordagens quimiométricas são introduzidas para a detecção de amostras não conformes em meio a misturas de biodiesel-diesel, pretendendo fornecer uma ferramenta útil para triagem de amostras antes da execução de análises laboriosas. Para este fim, aplicou-se a iPCA aos espectros de RMN ^{1}H (após redução da matriz em *buckets* [17]) centrados na média utilizando os intervalos mostrados na Tabela 5.

Inicialmente, um modelo com todas as amostras (que será comentado mais adiante) foi construído. As amostras reprovadas neste modelo (18 amostras), ou seja, aquelas encontradas fora dos elipsoides com 95% de confiança para os agrupamentos das classes, em cada intervalo estudado por iPCA e as demais amostras reprovadas de acordo com os limites adotados na legislação Brasileira (40 amostras) [76], foram selecionadas para compor o conjunto de validação (58 amostras), sendo as demais usadas para compor o conjunto de treinamento (42 amostras). A legislação citada especifica para o controle de qualidade das misturas aqui estudadas, alguns parâmetros físico-químicos, tais como: teor de biodiesel, massa específica (densidade) a 20 °C, temperatura de destilação de 50 e 85% do volume da amostra, índice de cetano, teor de enxofre e ponto de fulgor (Tabela 3).

As amostras do conjunto de treinamento (contendo as duas classes) foram utilizadas para construir um novo modelo iPCA. A seguir, as amostras do conjunto de validação foram centradas na média das amostras do conjunto de treinamento e projetadas nos subespaço gerado pelos pesos do modelo iPCA, obtendo-se a predição dos seus escores. As Figuras 21 a 24 mostram as elipses com 95% de confiança para todos os intervalos, juntamente com os escores preditos para as amostras reprovadas.

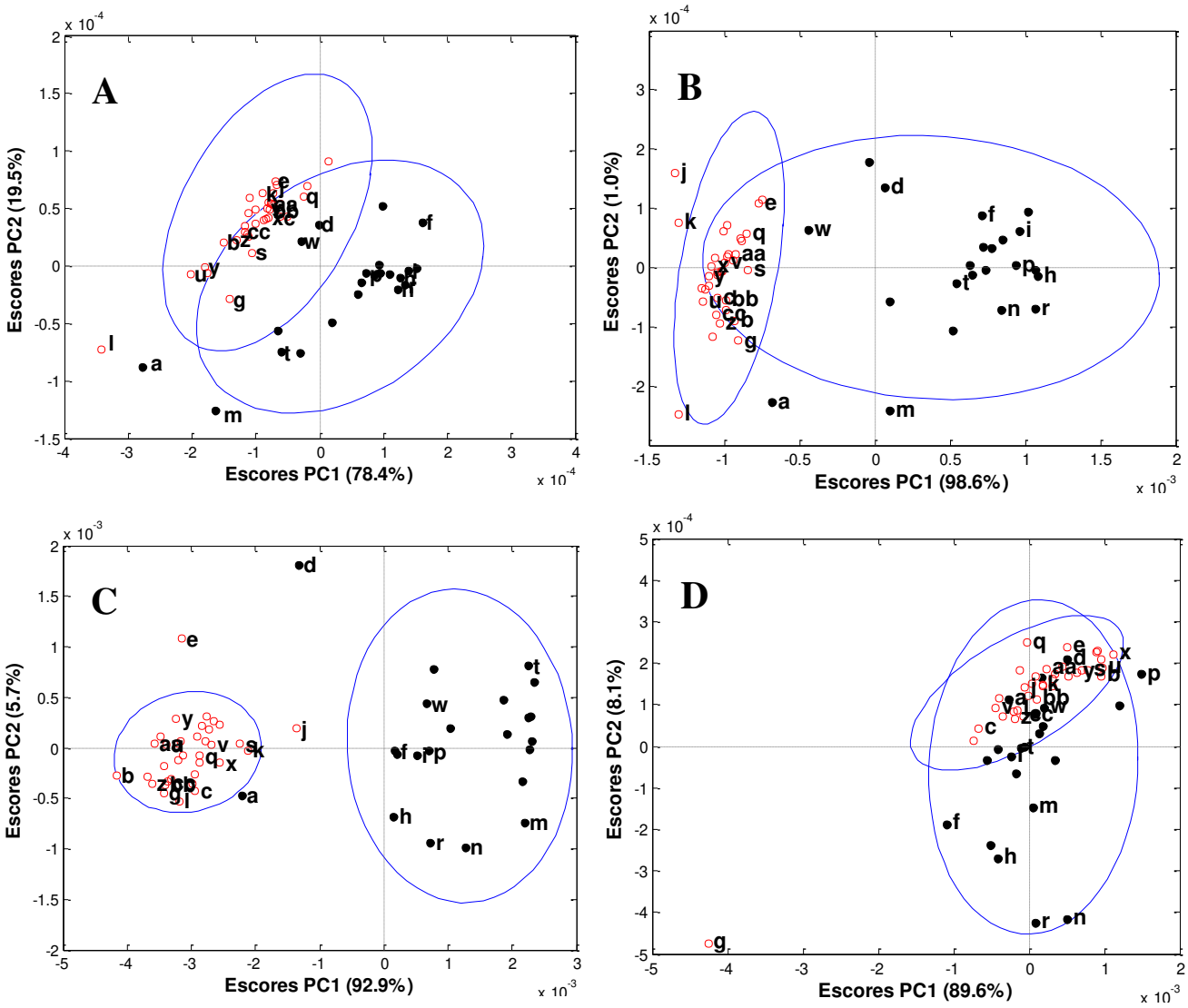

Figura 21: iPCA dos espectros de RMN ^{1}H das amostras não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. Gráficos de escores para (A) intervalo 1; (B) intervalo 2; (C) intervalo 3 e (D) intervalo 4. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

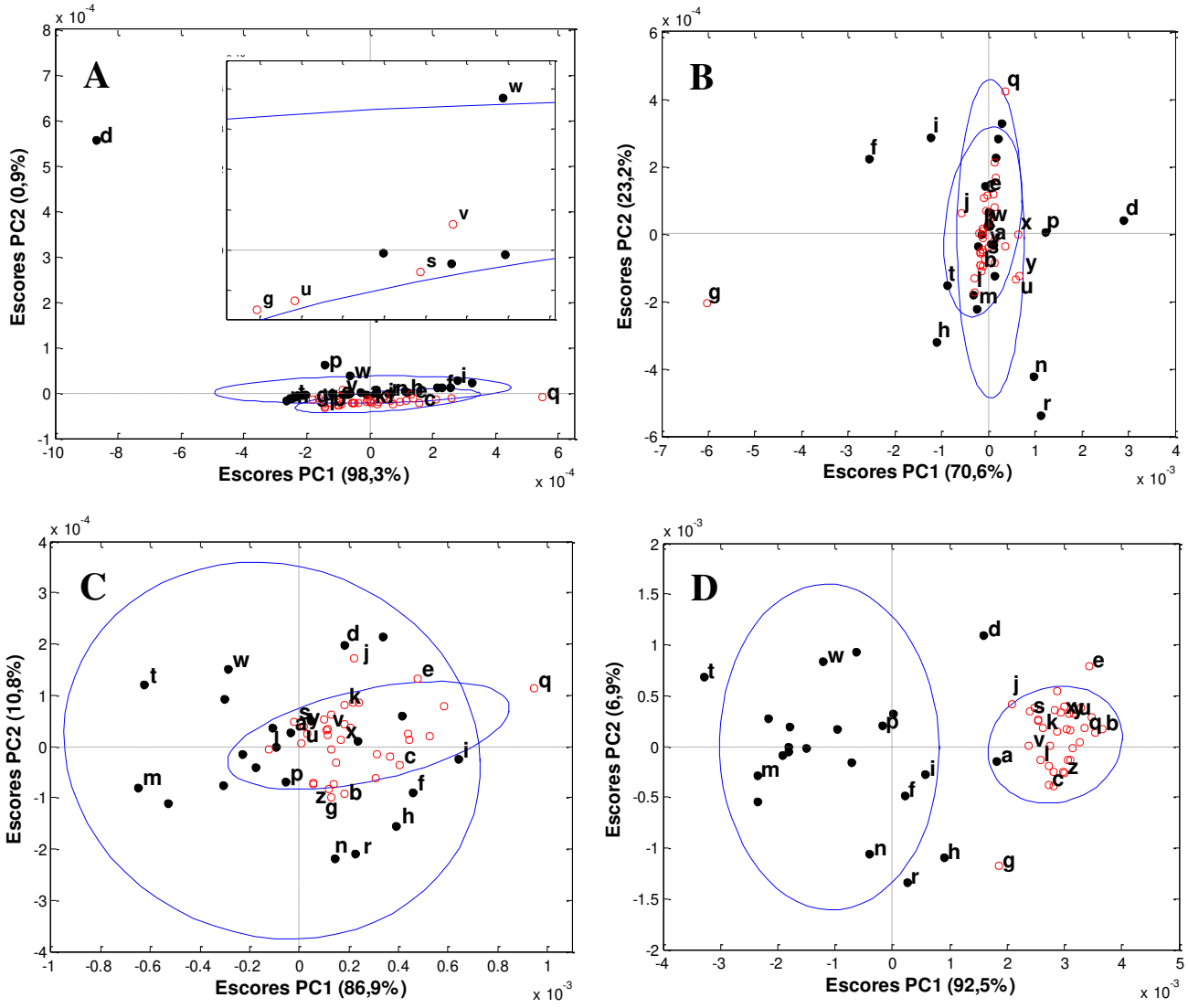

Figura 22: iPCA dos espectros de RMN ^{1}H das amostras não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. Gráficos de escores para (A) intervalo 5; (B) intervalo 6; (C) intervalo 7 e (D) intervalo 8. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

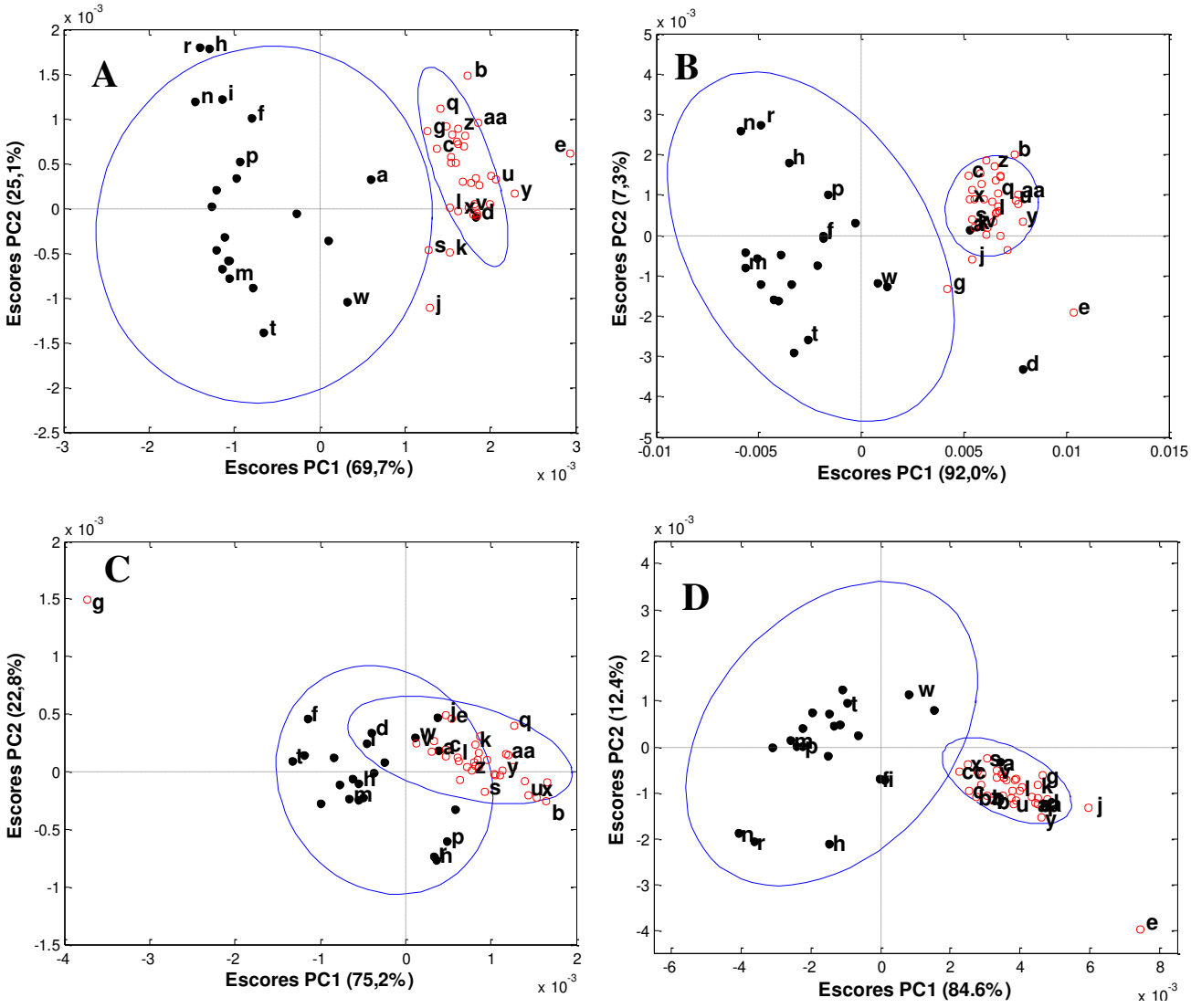

Figura 23: iPCA dos espectros de RMN ^{1}H das amostras não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. Gráficos de escores para (A) intervalo 9; (B) intervalo 10; (C) intervalo 11 e (D) intervalo 12. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

Os intervalos evidenciam diferentes amostras fora da especificação, sendo a maioria detectada pelas informações presentes na segunda componente principal (PC2). As amostras discrepantes apontadas pelo modelo são identificadas pelas letras: “a”, “b”, “c”, “d”, “e”, “f”, “g”, “h”, “i”, “j”, “k”, “l”, “m”, “n”, “p”, “q”, “r”, “s”, “t”, “u”, “v”, “w”, “x”, “y”, “z”, “aa”, “bb” e “cc”. Dentre estas amostras, apenas cinco

("c", "e", "j", "l" e "r") não são amostras reprovadas pelos parâmetros físico-químicos na Tabela 3, sendo consideradas não conformes pelo modelo de análise exploratória, portanto, com base apenas no padrão espectral estudado. No entanto, é importante citar que os parâmetros mostrados na Tabela 3 são apenas alguns daqueles avaliados para a qualidade das misturas de biodiesel-diesel, e assim não se descarta a possibilidade das amostras detectadas estarem fora da especificação no que diz respeito a algum outro parâmetro não apresentado.

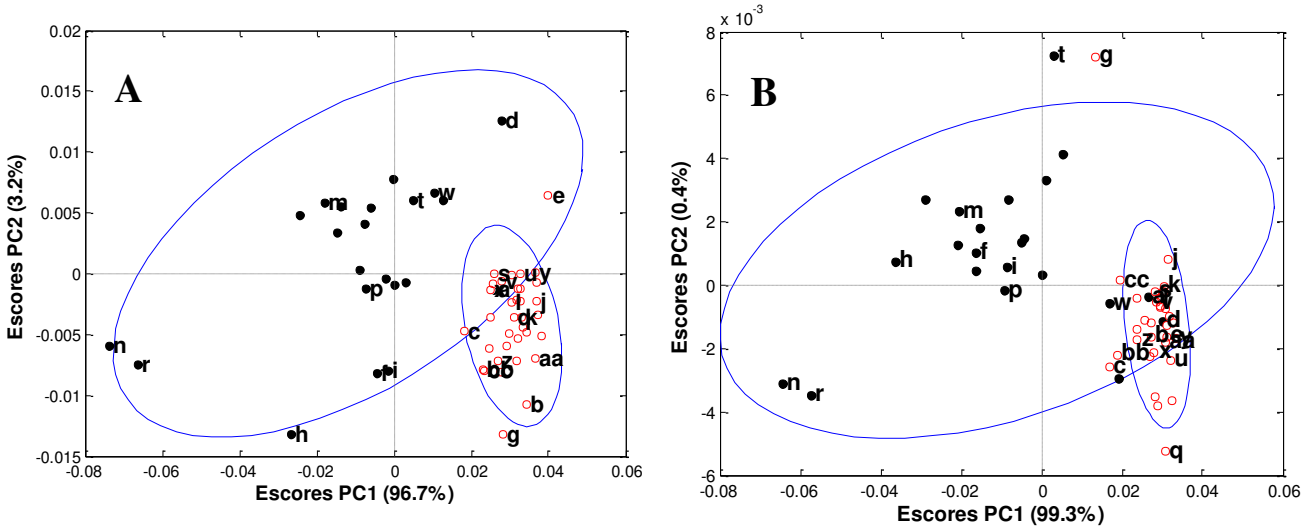

Figura 24: iPCA dos espectros de RMN ^{1}H das amostras não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. Gráficos de escores para (A) intervalo 13 e (B) intervalo 14. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

Além do exposto acima, vale ressaltar que no contexto da metodologia aqui apresentada, as amostras "c", "e", "j", "l" e "r" devem fazer parte do conjunto de validação, pois a presença das mesmas no conjunto de treinamento tende a proporcionar a obtenção de elipses de confiança com eixos maiores, sendo, portanto, "menos exigentes" na avaliação de misturas fora da especificação, e assim limitando o controle de qualidade proposto. Adicionalmente, se uma mistura, como por exemplo,

a amostra “e” (que se destaca bastante das demais em alguns intervalos) fizesse parte do conjunto de treinamento, seria necessária uma componente principal exclusivamente para explicar a característica desta amostra, afetando o desempenho para a descrição das demais.

A

Figura 25A indica, através dos espectros de RMN ^{1}H após o pré-tratamento por *bucketing* otimizado, que as amostras “c”, “e”, “j”, “l” e “r” apresentam um perfil geral do espectro bastante similar entre si, bem como quando comparadas à média das amostras conformes (amostras não reprovadas na Tabela 3), não apresentando picos distintos. Contudo, a análise das Figuras 25B - F mostra que as misturas, nos intervalos em que são detectadas, apresentam espectros com intensidades diferentes daquele da média das amostras não reprovadas, o que pode justificar as suas detecções, mesmo sendo estas, amostras não reprovadas pelos parâmetros físico-químicos. Provavelmente, tais diferenças são somente observadas devido ao uso do estudo dos espectros em intervalos, o que permite que algumas diferenças sutis entre as amostras sejam evidenciadas, além de possibilitar uma análise menos complexa do que aquela usando o espectro inteiro.

De fato, o modelo PCA, apresentado na Figura 26, construído a partir dos espectros de RMN ^{1}H inteiros utilizando o mesmo conjunto de treinamento do espectro iPCA acima, mostra-se mais limitado para a detecção das amostras não conformes no conjunto de validação, evidenciando apenas as amostras: “c”, “d”, “e”, “g”, “j”, “n”, “r”, “y”, “bb” e “cc”, além de três amostras ainda não rotuladas. As quatro primeiras componentes principais deste modelo PCA explicam 99,7% da variância total e apresentam através dos pesos das variáveis o envolvimento de vários picos na discriminação das amostras. O provável mascaramento de sinais discriminatórios pequenos, quando a análise do espectro inteiro é realizada, deve ser o

responsável pelo desempenho inferior do modelo PCA frente ao modelo iPCA, apesar deste último desconsiderar a correlação entre as variáveis de diferentes intervalos.

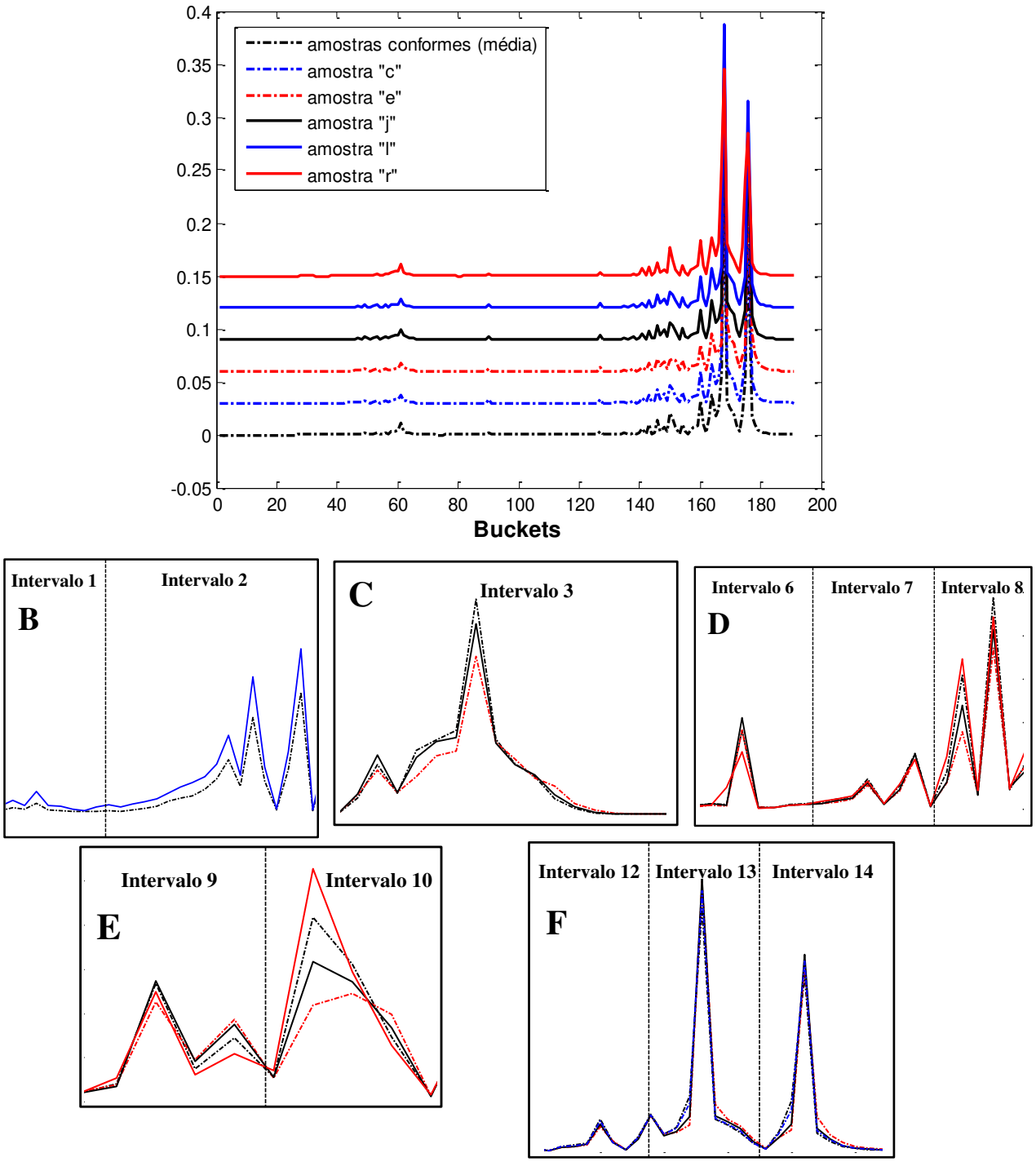

Figura 25: Comparação entre os espectros de RMN ^{1}H (após a transformação com o *bucketing* otimizado) das amostras “c” (verde), “e” (amarelo), “j” (preto), “l” (azul) e “r” (vermelho) com o espectro médio das amostras não reprovadas (ciano). Em A os espectros encontram-se deslocados nas ordenadas. Os intervalos são mostrados como

regiões ampliadas contendo o(s) espectro(s) da(s) amostra(s) detectada(s) e o espectro médio das amostras não reprovadas.

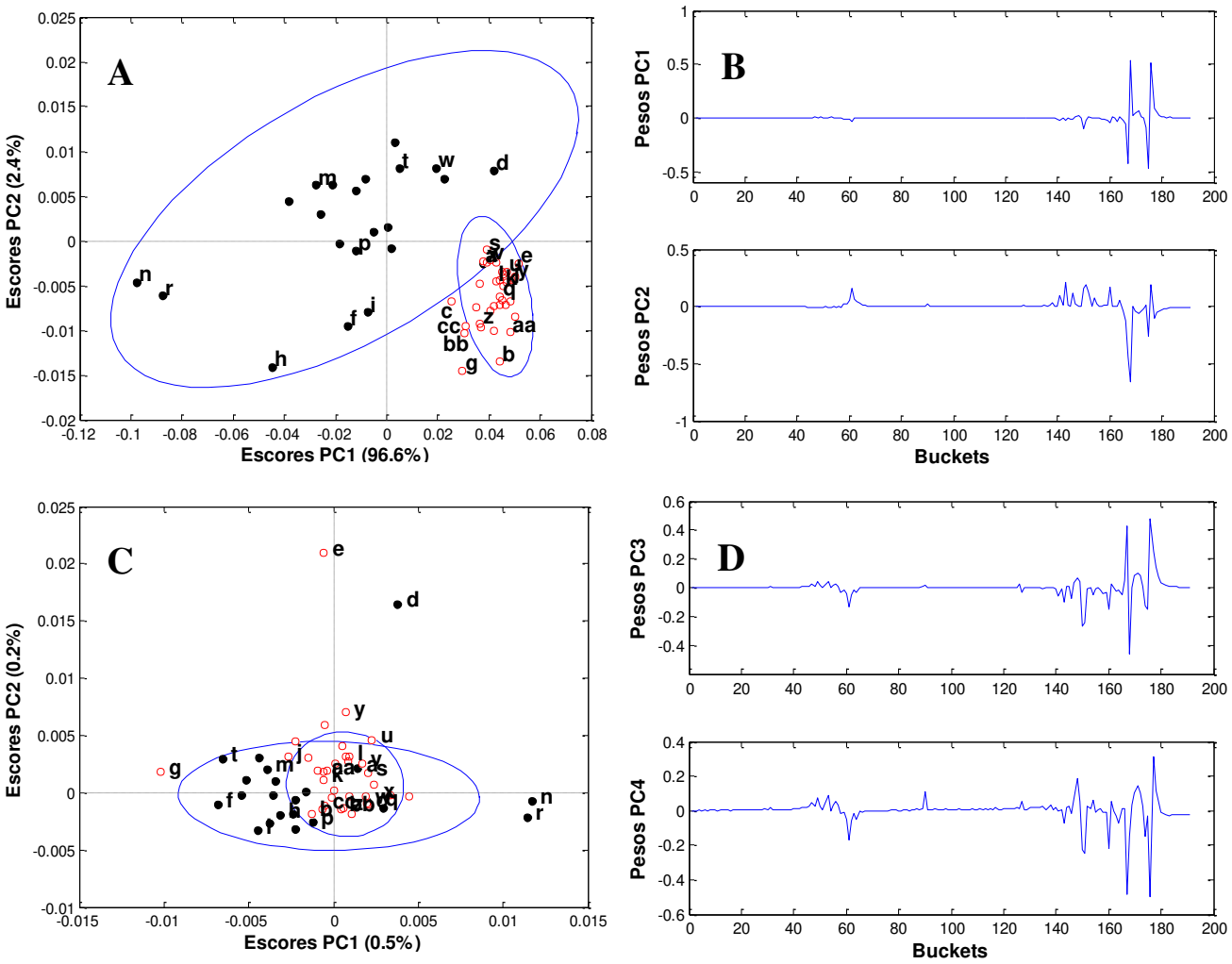

Figura 26: PCA dos espectros de RMN ^{1}H. A e B, gráficos de escores e de pesos – PC1 *versus* PC2; C e D, gráficos de escores e de pesos – PC3 *versus* PC4. Amostras metropolitanas (conjunto de validação) são identificadas por pontos pretos, enquanto as amostras não metropolitanas (conjunto de validação) são identificadas por círculos vermelhos.

Além do modelo iPCA incluindo as duas classes de amostras, modelos para cada classe separada também podem ser úteis na identificação de amostras problemáticas, como mostrado nas Figuras 27 a 33. Os modelos foram construídos a partir das amostras não reprovadas de cada classe, sendo as amostras reprovadas

centradas na média das primeiras e projetadas no subespaço gerado pelos pesos nos modelos. Deste modo, modelos para todos os intervalos e para as duas classes foram obtidos.

Observa-se que os modelos para cada classe detectam todas as amostras apontadas pelos modelos com as duas classes, além de evidenciar outras amostras, por exemplo, para as misturas metropolitanas, as amostras: 62 (“dd”) nos intervalos 4 (Figura 28C) e 13 (Figura 33A); 98 (“ee”) nos intervalos 4 (Figura 28C) e 11 (Figura 32A); 68 (“ff”) no intervalo 5 (Figura 29A); 11 (“gg”) no intervalo 6 (Figura 29C); 80 (“hh”) no intervalo 10 (Figura 31C); 71 (“ii”) no intervalo 10 (Figura 31C); e 31 (“jj”) nos intervalos 12 (Figura 32C) e 14 (Figura 33C).

Dentre as amostras citadas, aquelas “dd”, “ee”, “ff”, “gg” e “hh” são detectadas apenas por este modelo iPCA, não sendo reprovadas pelos parâmetros físico-químicos na Tabela 3. Assim, é provável que a detecção destas amostras esteja sustentada em algum parâmetro não mostrado na Tabela 3, do mesmo modo que foi comentado acima, sobre as amostras “c”, “e”, “j”, “l” e “r”.

As amostras “ii” e “jj” já haviam sido reprovadas pelos parâmetros citados devido a distintos fatores, a saber, baixos, índice de cetano e teor de biodiesel (Tabela 3), respectivamente, o que pode explicar os diferentes intervalos em que são detectadas. Portanto, a adição destas amostras dentre as reprovadas significa um ganho no que diz respeito ao uso de modelos para cada classe separada.

Para a classe não metropolitana, detectaram-se misturas também já reprovadas pelos parâmetros na Tabela 3, como as amostras: 33 (“kk”) no intervalo 7 (Figura 30B); 61 (“ll”) nos intervalos 7 (Figura 30B), 9 (Figura 31B), 10 (Figura 31D) e 13 (Figura 33B); 89 (“mm”) nos intervalos 7 (Figura 30B) e 13 (Figura 33B); e 95 (“nn”) nos intervalos 7 (Figura 30B) e 10 (Figura 31D). A detecção destas amostras também aponta para uma melhoria devido ao uso dos modelos aqui comentados.

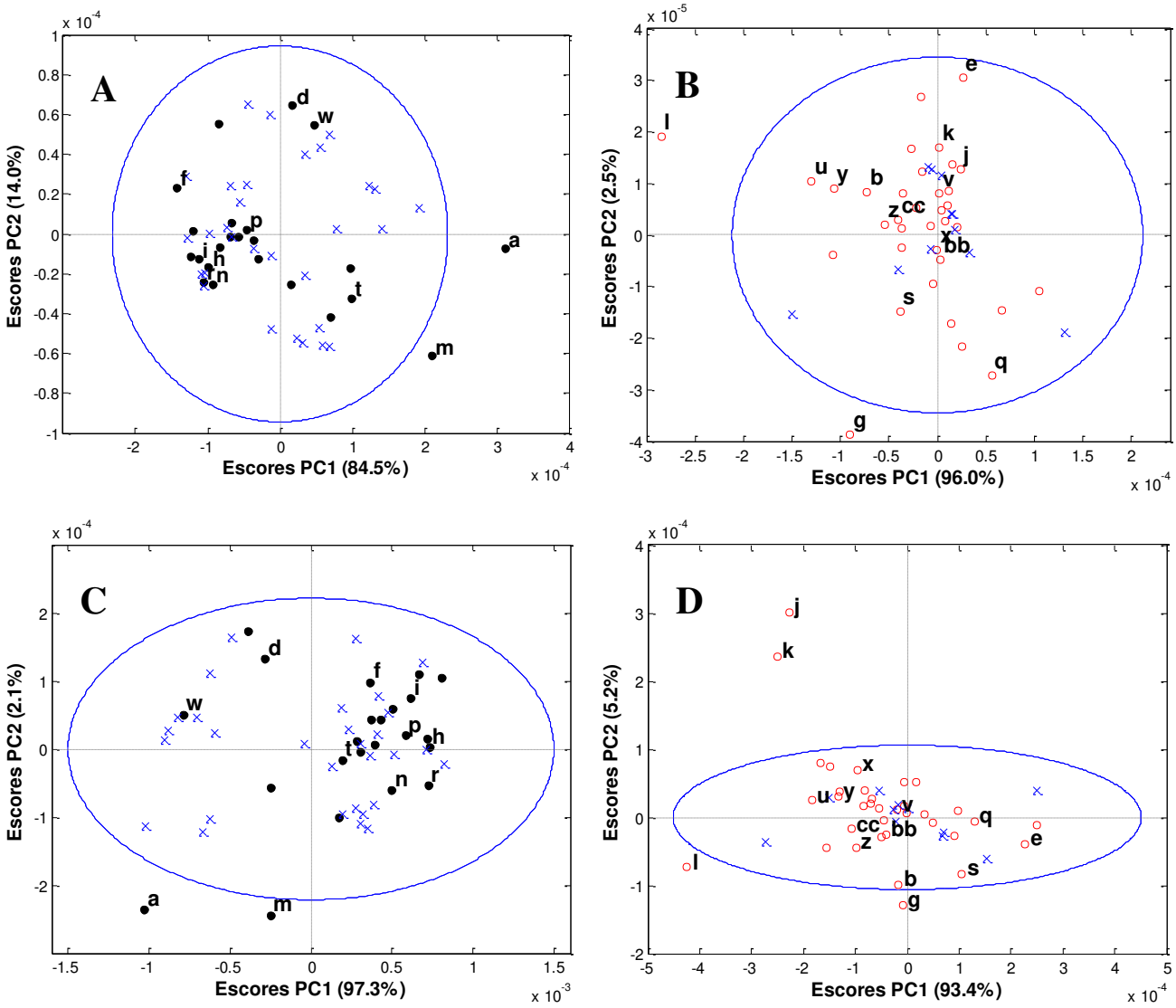

Figura 27: iPCA dos espectros de RMN ^{1}H para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 1) e C (intervalo 2): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 1) e D (intervalo 2): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

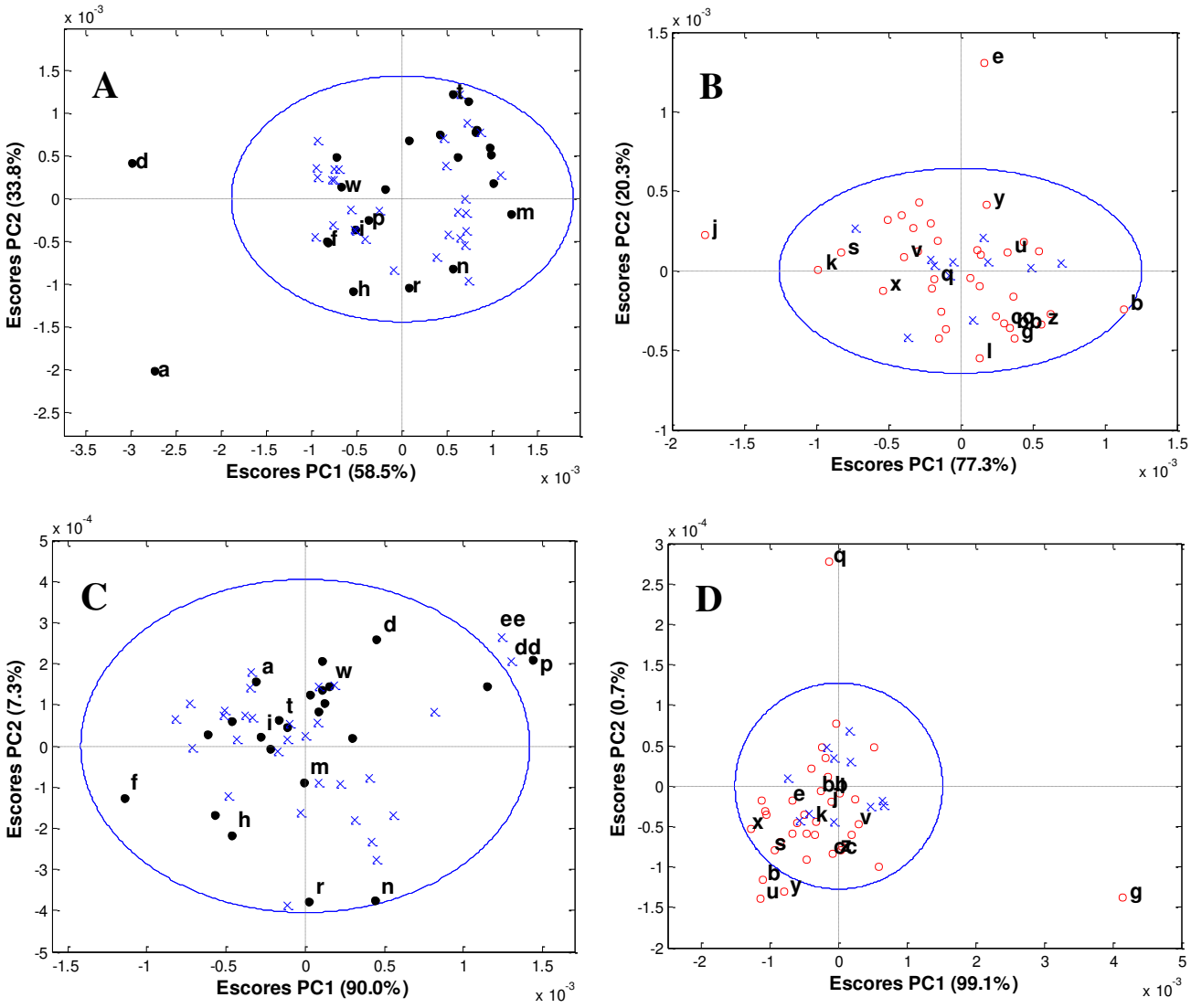

Figura 28: iPCA dos espectros de RMN ^{1}H para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 3) e C (intervalo 4): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 3) e D (intervalo 4): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

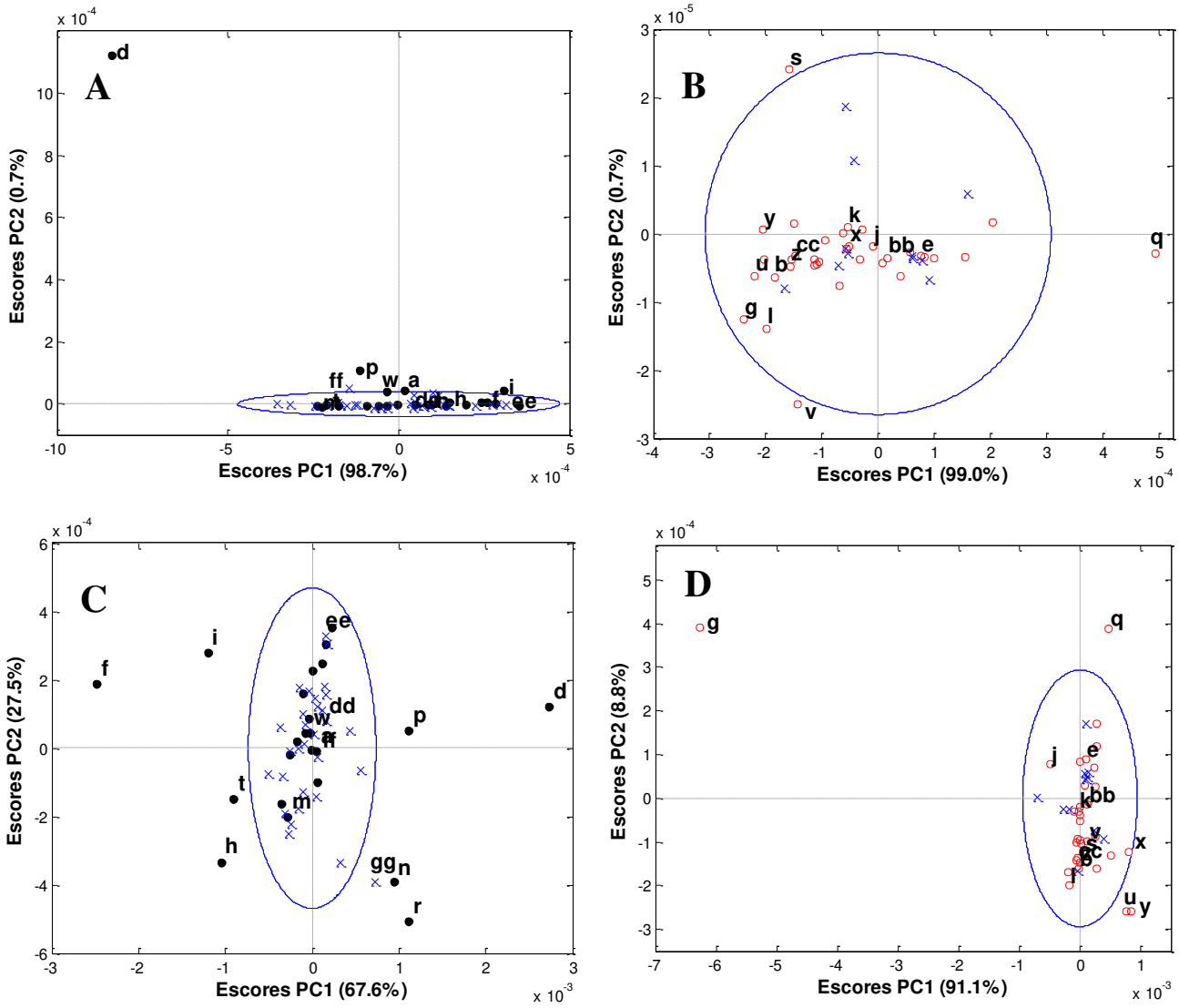

Figura 29: iPCA dos espectros de RMN ^{1}H para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 5) e C (intervalo 6): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 5) e D (intervalo 6): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

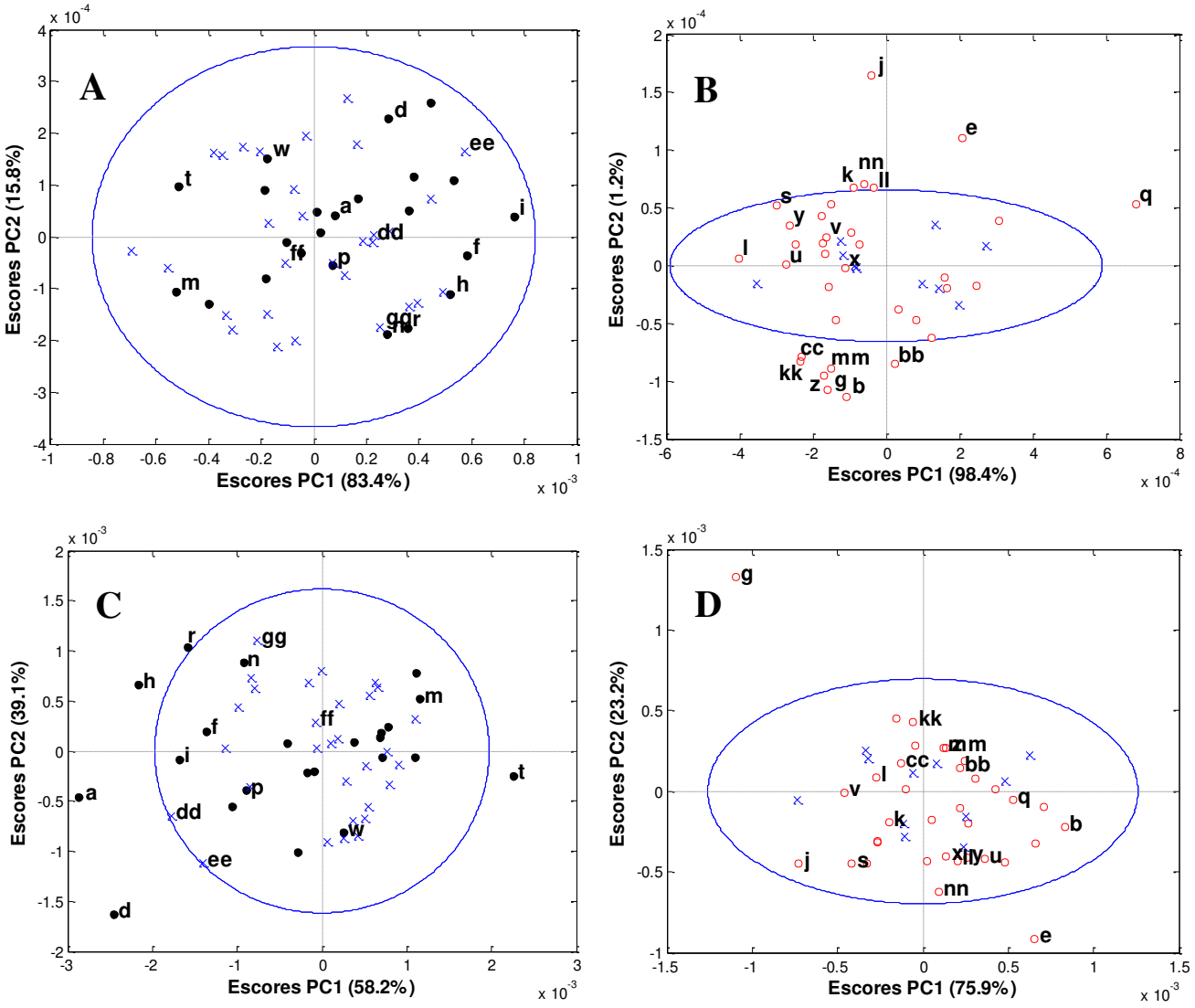

Figura 30: iPCA dos espectros de RMN ^{1}H para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 7) e C (intervalo 8): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 7) e D (intervalo 8): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

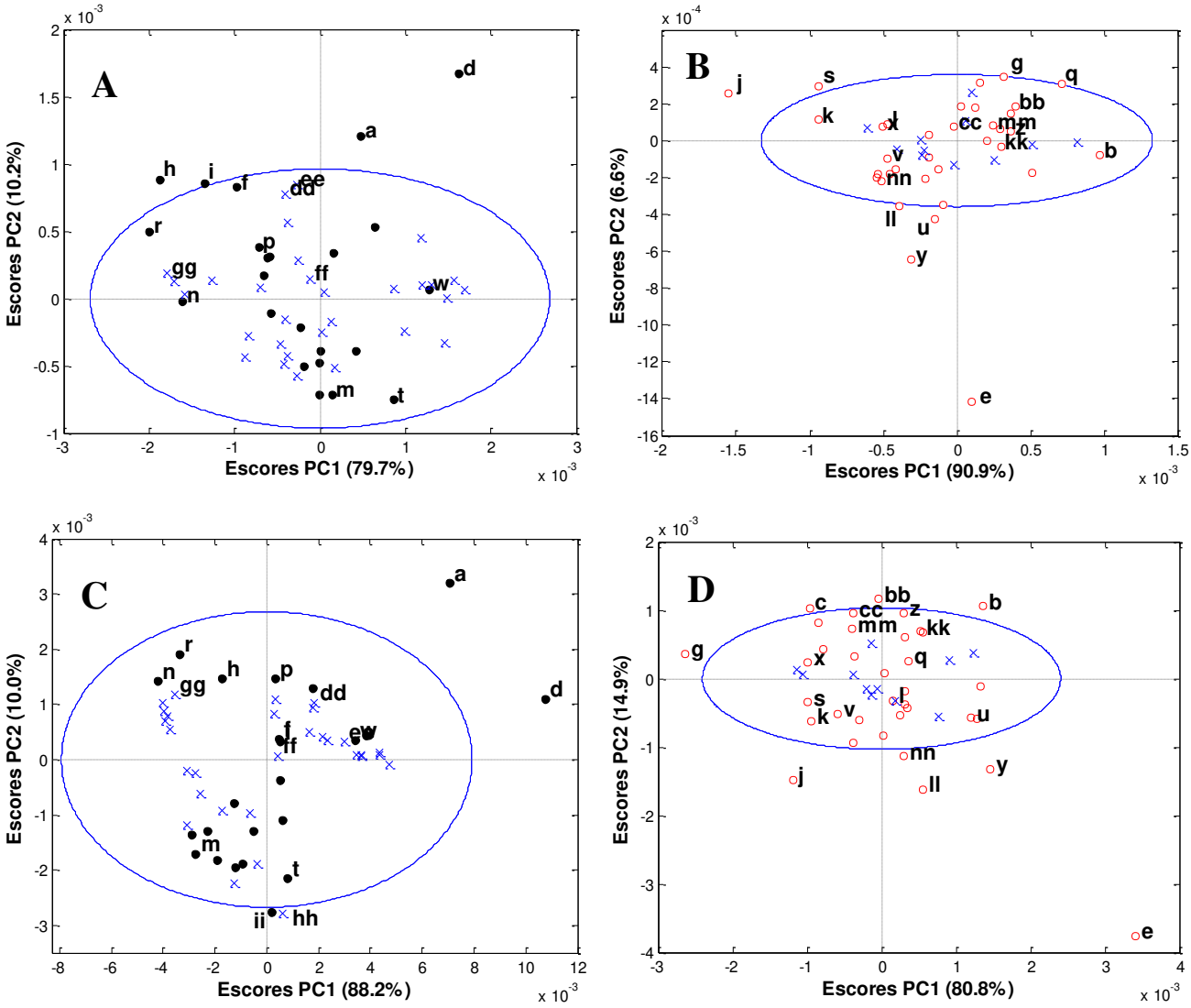

Figura 31: iPCA dos espectros de RMN ^{1}H para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 9) e C (intervalo 10): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 9) e D (intervalo 10): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

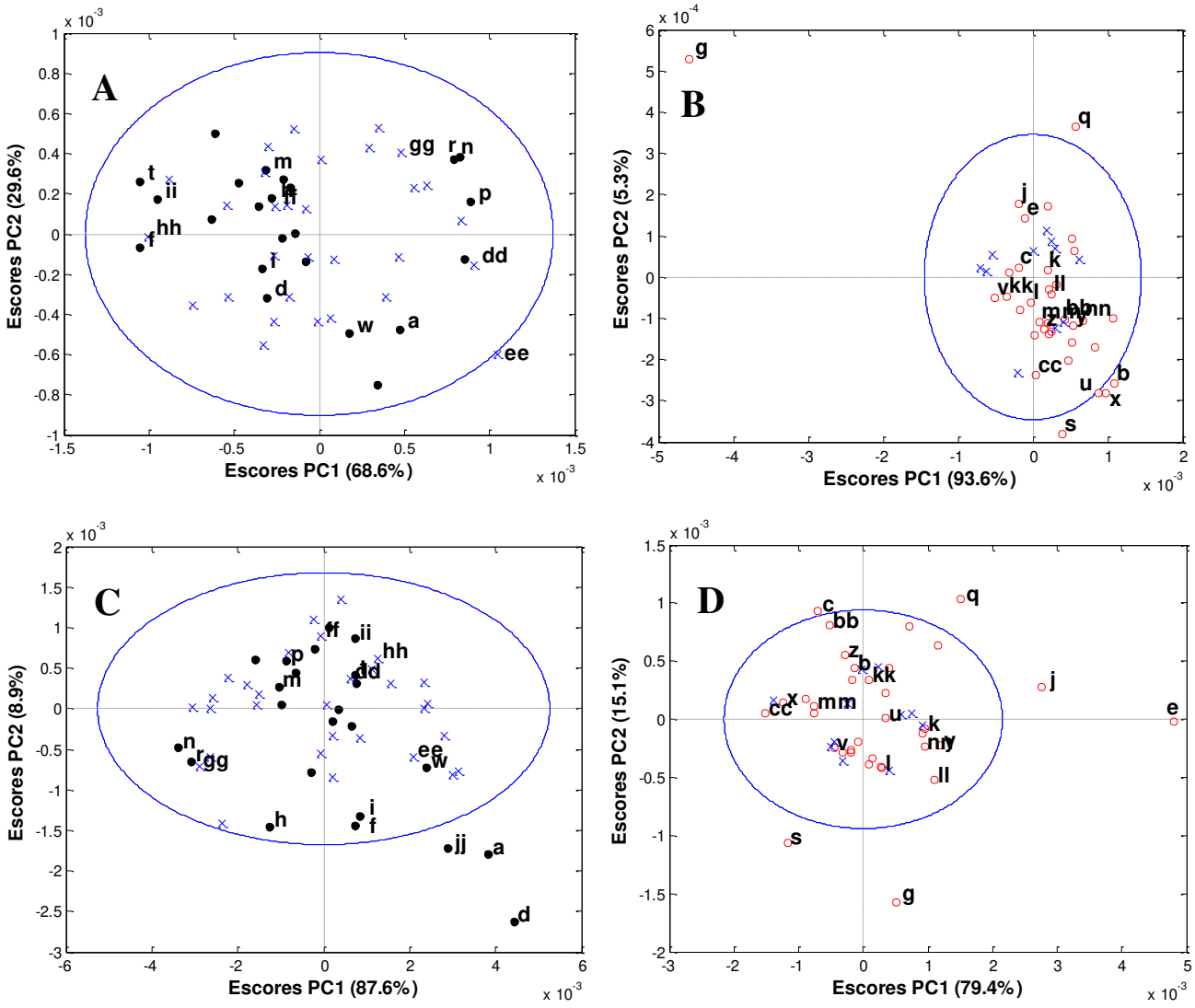

Figura 32: iPCA dos espectros de RMN ^{1}H para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 11) e C (intervalo 12): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 11) e D (intervalo 12): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

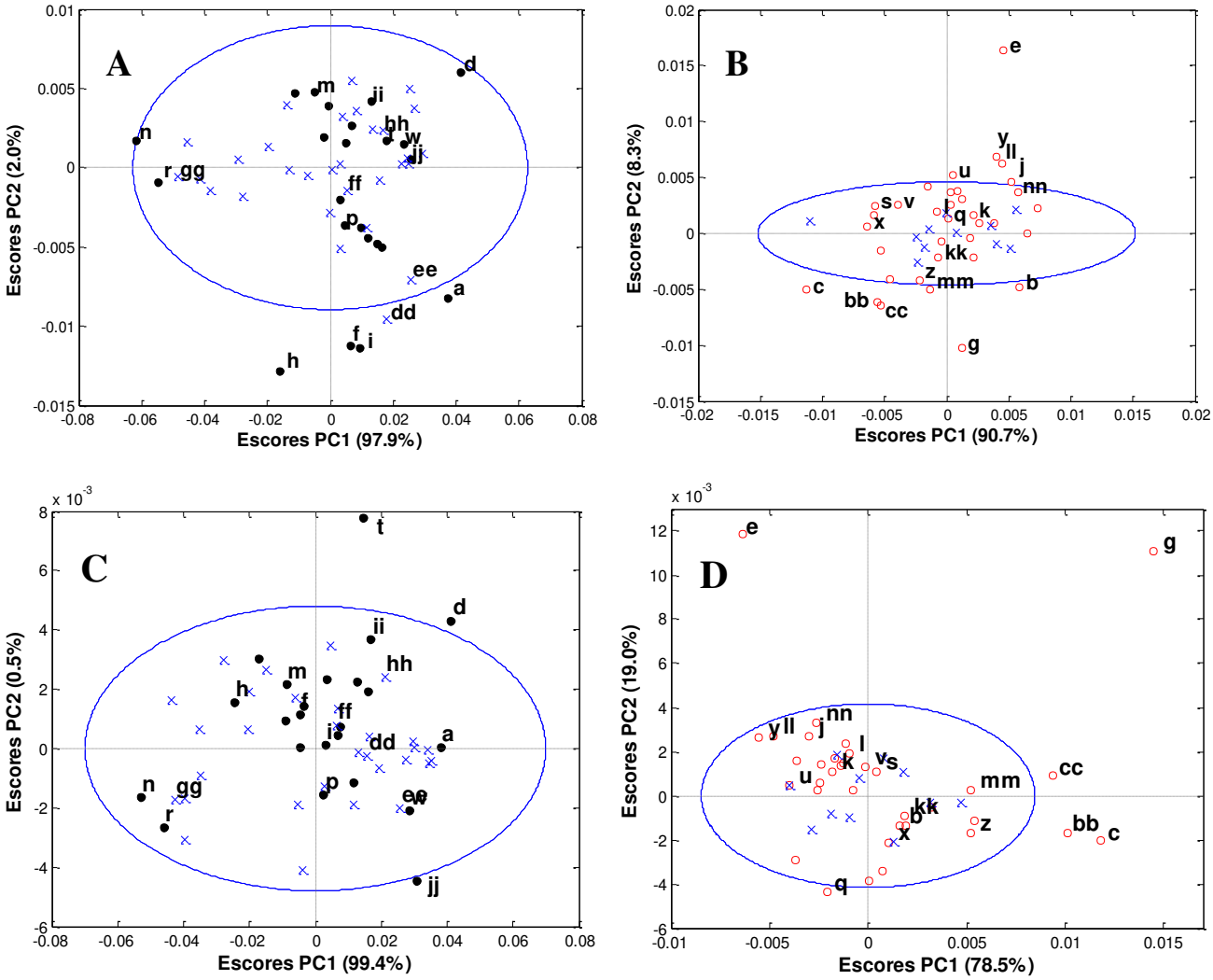

Figura 33: iPCA dos espectros de RMN ^{1}H para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 13) e C (intervalo 14): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 13) e D (intervalo 14): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

Nos modelos usando as duas classes juntas, a máxima variação, capturada pela primeira componente principal, corresponde às diferenças entre as classes (praticamente em todos os intervalos), assim para os modelos com uma classe, esta variação é removida e as componentes principais tendem a descrever melhor as relações apenas entre as amostras do mesmo tipo. Isto equivale mais ou menos à

remoção de *outliers* do conjunto, uma vez que as amostras de uma classe são bastante distintas daquelas da outra classe. Em outras palavras, há uma mudança de foco de "diferença entre classes" para "diferença dentro das classes", sendo os resultados devido a esta mudança bastante favoráveis.

Entretanto, em algumas situações a presença das duas classes pode ser útil para a detecção de algumas amostras não conformes específicas e para permitir inferências sobre as características químicas das amostras, uma vez que a exploração de modelos com uma variabilidade química maior tende a favorecer as interpretações. Dessa forma, neste trabalho, recomenda-se a investigação das amostras tanto num modelo iPCA com as duas classes quanto nos modelos iPCA para as classes isoladas, uma vez que a variabilidade dos modelos pode apontar diferentes amostras discrepantes, atuando os modelos como análises complementares.

Apesar das vantagens mostradas pelos modelos iPCA, uma limitação é observada no que diz respeito à detecção das amostras fora da especificação para o teor de biodiesel. O limite permitido pela legislação, à época da coleta das amostras estudadas, era de 4,0 ± 0,5% em volume de biodiesel na mistura, no entanto, observa-se na Tabela 3 que praticamente a metade das amostras encontra-se fora da especificação. Mesmo assim, nos modelos, muitas amostras estão dentro das elipses de confiança e, portanto, são erradamente consideradas "conformes". Apenas as amostras com teores muito acima (por exemplo, as amostras "f", "g", "h" e "i") ou muito abaixo (por exemplo, a amostra "d") do limite são claramente reprovadas.

Isto ocorre, provavelmente, dado o baixo número de amostras "conformes" utilizadas para a construção dos modelos, o que limita a variabilidade dos mesmos para detectar este parâmetro. Este ponto fraco pode ser contornado utilizando um maior número de amostras "conformes" ou ainda diminuindo a tolerância das elipses, pelo uso de um nível de confiança menor.

Pode-se ainda argumentar que para o teor de biodiesel, o padrão espectral pode não ser adequado devido a uma pequena diferença entre os espectros das amostras "conformes" e "não conformes", não permitindo a discriminação desejada. Em outras palavras, a diferença espectral entre amostras com teor de biodiesel de 3,5% (conforme) e 3,4% (não conforme), por exemplo, pode não ser grande o suficiente para que esta última seja detectada. Na Tabela 3 pode ser visto que dezenove amostras fora da especificação para o teor de biodiesel não foram reprovadas pela metodologia proposta. Este número baixa para apenas três amostras, se em vez de um limite de 4,0 ± 0,5% for utilizado 4,0 ± 1,0%. Neste último caso, é provável que uma diferenciação maior no padrão espectral possibilite a melhoria citada.

Finalmente, vale ressaltar que a divisão dos intervalos realizada para a construção dos modelos iPCA pode sofrer mudanças de acordo com o conjunto de dados analisado, pois não é possível prever que tipo de adulterações as amostras podem sofrer. Na análise aqui realizada, as regiões entre 5,6 – 6,1 ppm e 4,2 – 6,0 ppm não foram investigadas por não possuírem sinais, mas futuras análises podem ser estendidas para as mesmas, no intuito de tornar os modelos mais abrangentes e capazes de detectar características inesperadas nas amostras.

3.5.2.1. Espectros de RMN ^{1}H – características químicas das classes

Um modelo iPCA foi construído utilizando as cem amostras de biodiesel-diesel disponíveis. Os resultados deste modelo são mostrados nas Figuras 34 a 40, onde as duas primeiras componentes principais são analisadas. Observa-se que quase todos os intervalos indicam separações entre as amostras metropolitanas e não metropolitanas e evidenciam diferentes amostras fora da especificação, ao longo de cada "janela" de deslocamentos químicos em observação.

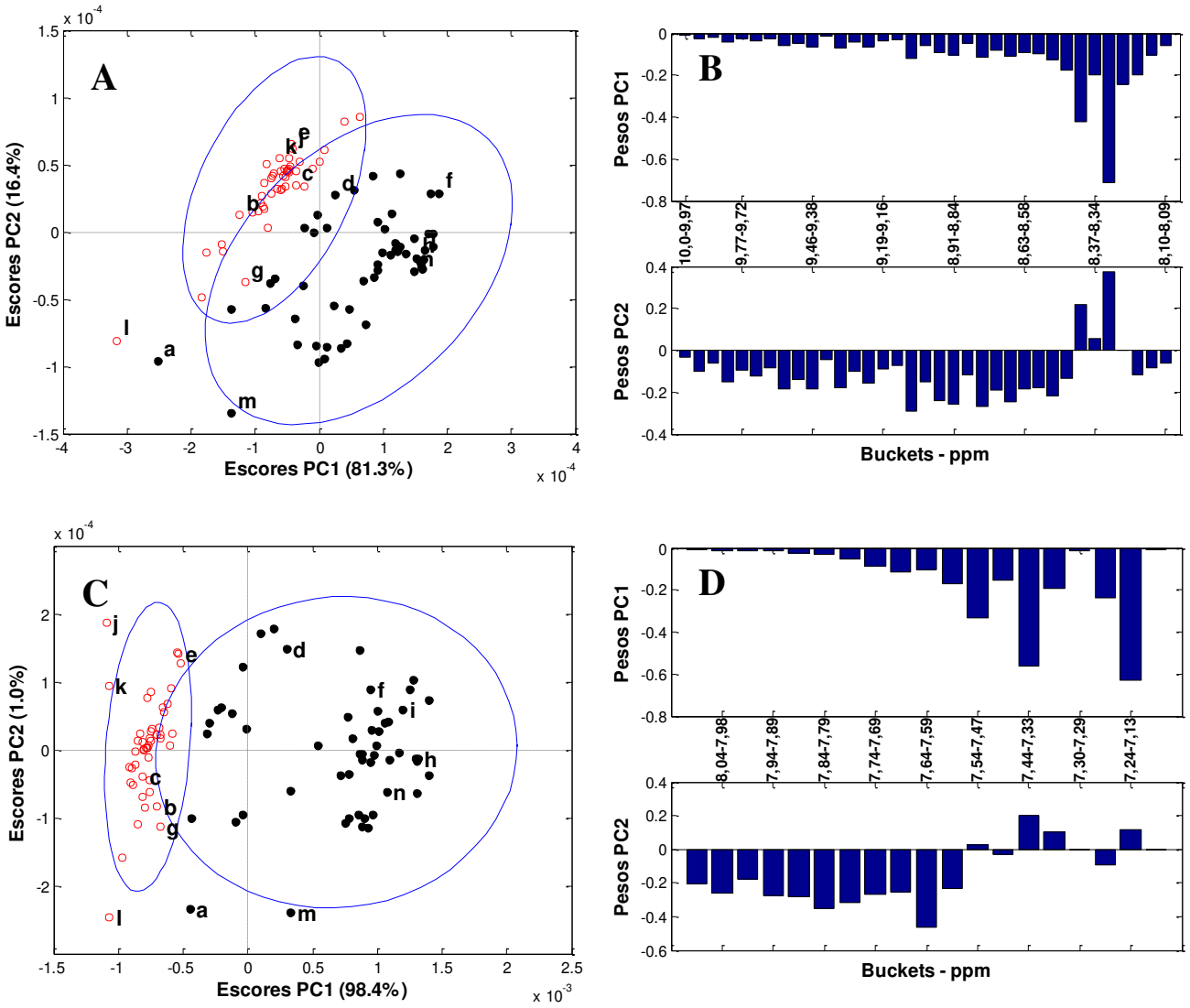

Figura 34: iPCA dos espectros de RMN ^{1}H. A e B, gráficos de escores e de pesos relativos ao intervalo 1; C e D, gráficos de escores e de pesos relativos ao intervalo 2. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

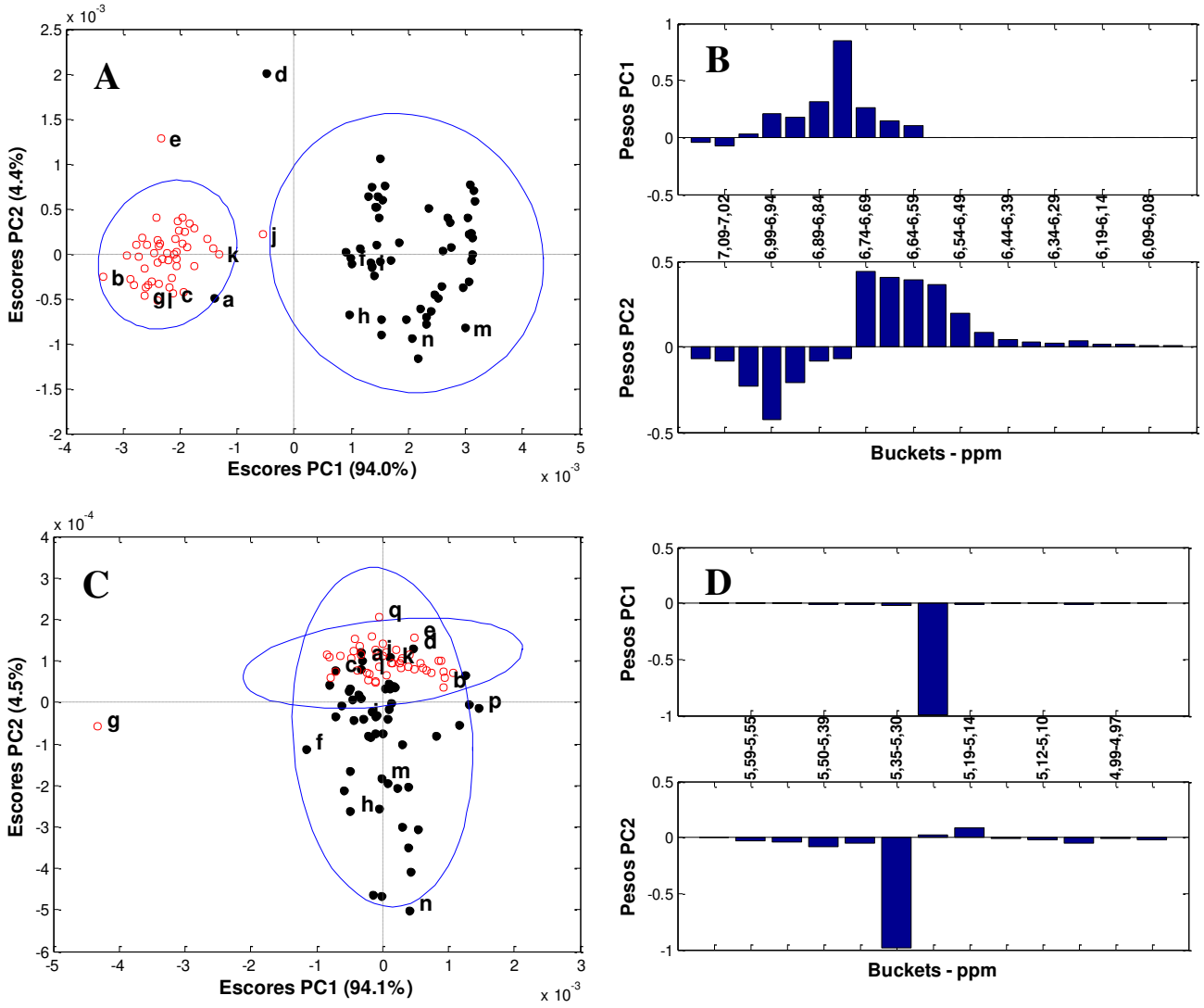

Figura 35: iPCA dos espectros de RMN ^{1}H. A e B, gráficos de escores e de pesos relativos ao intervalo 3; C e D, gráficos de escores e de pesos relativos ao intervalo 4. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

Comparando as amostras metropolitanas com as não metropolitanas através dos gráficos de escores e de pesos nas Figuras 34A, 34B, 34C e 34D (intervalos 1 e 2) infere-se que as amostras não metropolitanas têm o maior conteúdo de substâncias aromáticas polinucleares e os sinais em 7,24-7,13; 7,44-7,33; 7,54-7,47; 8,34-8,23 e 8,44-8,37 estão envolvidos na discriminação ao longo da primeira componente

principal, PC1, que conta com mais do que 80% de variância explicada em ambos os intervalos.

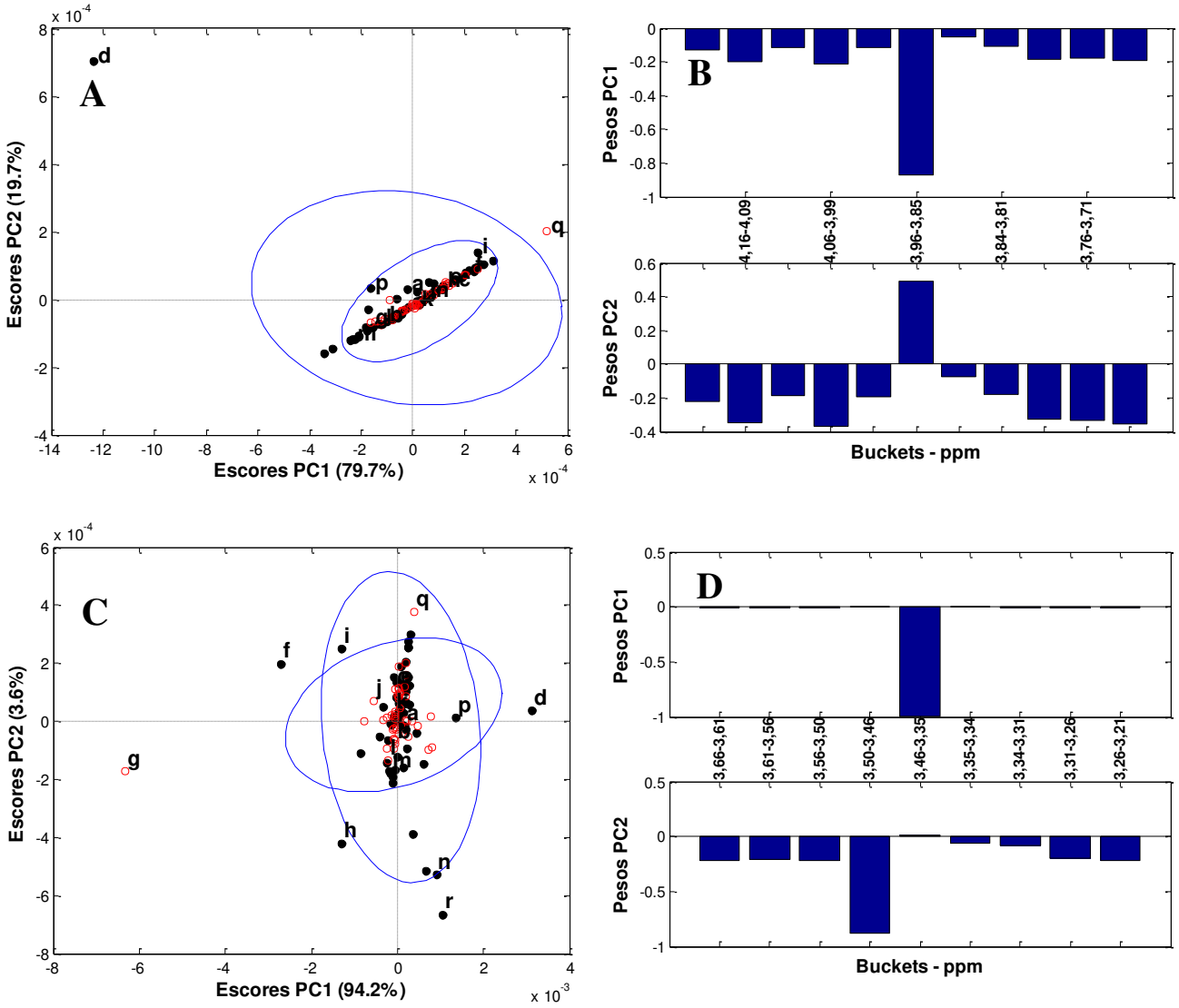

Figura 36: iPCA dos espectros de RMN ^{1}H. A e B, gráficos de escores e de pesos relativos ao intervalo 5; C e D, gráficos de escores e de pesos relativos ao intervalo 6. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

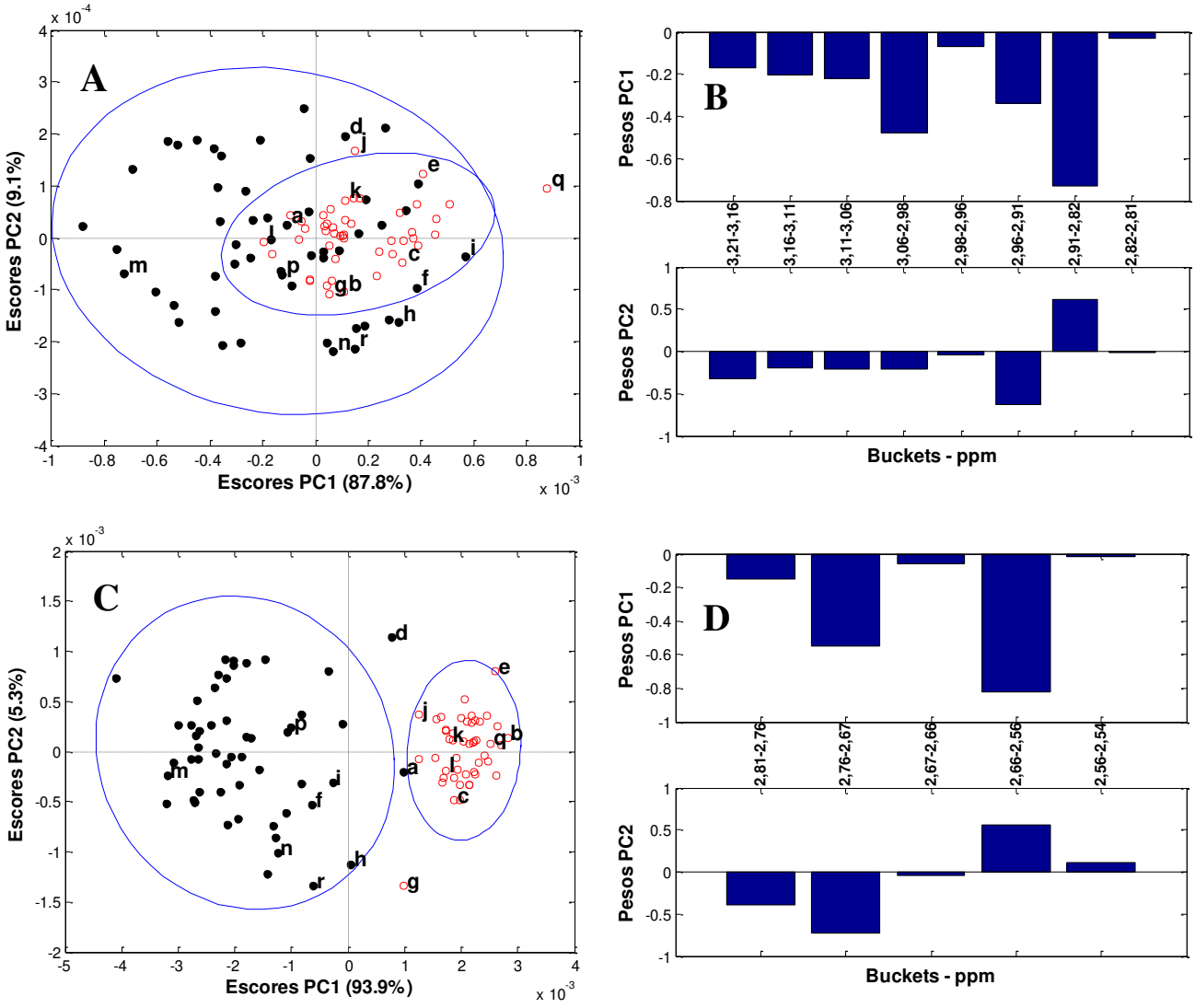

Figura 37: iPCA dos espectros de RMN ^{1}H. A e B, gráficos de escores e de pesos relativos ao intervalo 7; C e D, gráficos de escores e de pesos relativos ao intervalo 8. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

Por outro lado, o intervalo 3, nas Figuras 35A e 35B, indica que as amostras metropolitanas possuem um maior teor de aromáticos mononucleares, como mostrado pelos sinais em 6,85-6,74 ppm que possuem o maior valor de peso em PC1 (com 94% de variância explicada) e os sinais em 2,22-2,12 e 2,12-2,06 ppm (intervalo 10 na Figura 38) que se referem a ramificações em anéis aromáticos mononucleares. A distinção comentada pode ser decorrente do processo de hidrodessulfurização, o qual

as amostras metropolitanas são submetidas para diminuição do teor de enxofre. Neste processo alguns compostos aromáticos sulfurados, tais como, os polinucleares benzotiofeno e dibenzotiofeno são degradados em substâncias mais simples [74,75]. Ao final do processo, possivelmente, as amostras metropolitanas têm um menor conteúdo de compostos aromáticos polinucleares e maior conteúdo de aromáticos mononucleares.

Através dos gráficos de escores e pesos nas Figuras 35C e 35D (intervalo 4), pode ser notado que as amostras metropolitanas têm um maior conteúdo das substâncias responsáveis pelos deslocamentos químicos em 5,35-5,30 ppm (em PC2) referentes a prótons olefínicos em cadeias poli-insaturadas, o que pode ser explicado devido à mistura do diesel com produtos derivados do *cracking* do petróleo ou às diferentes fontes de matéria-prima para o biodiesel destas amostras, ou seja, os precursores do biodiesel das amostras metropolitanas podem ser mais instaurados levando o biodiesel a possuir uma característica similar. Observa-se ainda neste intervalo, em PC1, que a amostra identificada por "g" praticamente responde sozinha pelos deslocamentos químicos em 5,30-5,19 ppm.

As Figuras 38A e 38B mostram que as amostras não metropolitanas têm uma maior variabilidade em compostos aromáticos polinucleares ramificados devido à correlação destas amostras com os sinais em 2,41-2,36 e 2,36-2,26 ppm, em PC1 (com 74,2% de variância explicada), que se referem aos grupos ligados a estes núcleos aromáticos. Por outro lado, as Figuras 37C e 37D, relativas ao intervalo 8, apontam através dos sinais em 2,66-2,56 ppm ao longo de PC1 (com 93,9% de variância explicada) que as amostras metropolitanas possuem um maior conteúdo de ramificações em compostos aromáticos mononucleares, uma vez que estes sinais são decorrentes dos grupos $–CH_3$, $–CH_2–$, e –CH– ligados a estes aromáticos.

Os sinais em 3,46-3,35 e 3,50-3,46 ppm, referentes às metilas (ou etilas) ligadas às carbonilas das porções dos ésteres metílicos (ou etílicos) do biodiesel (intervalo 6,

Figuras 36C e 36D), demonstram que quase todas as amostras possuem conteúdos de biodiesel muito próximos, de acordo com os escores e pesos ao longo de PC1 (com 94,2% de variância explicada).

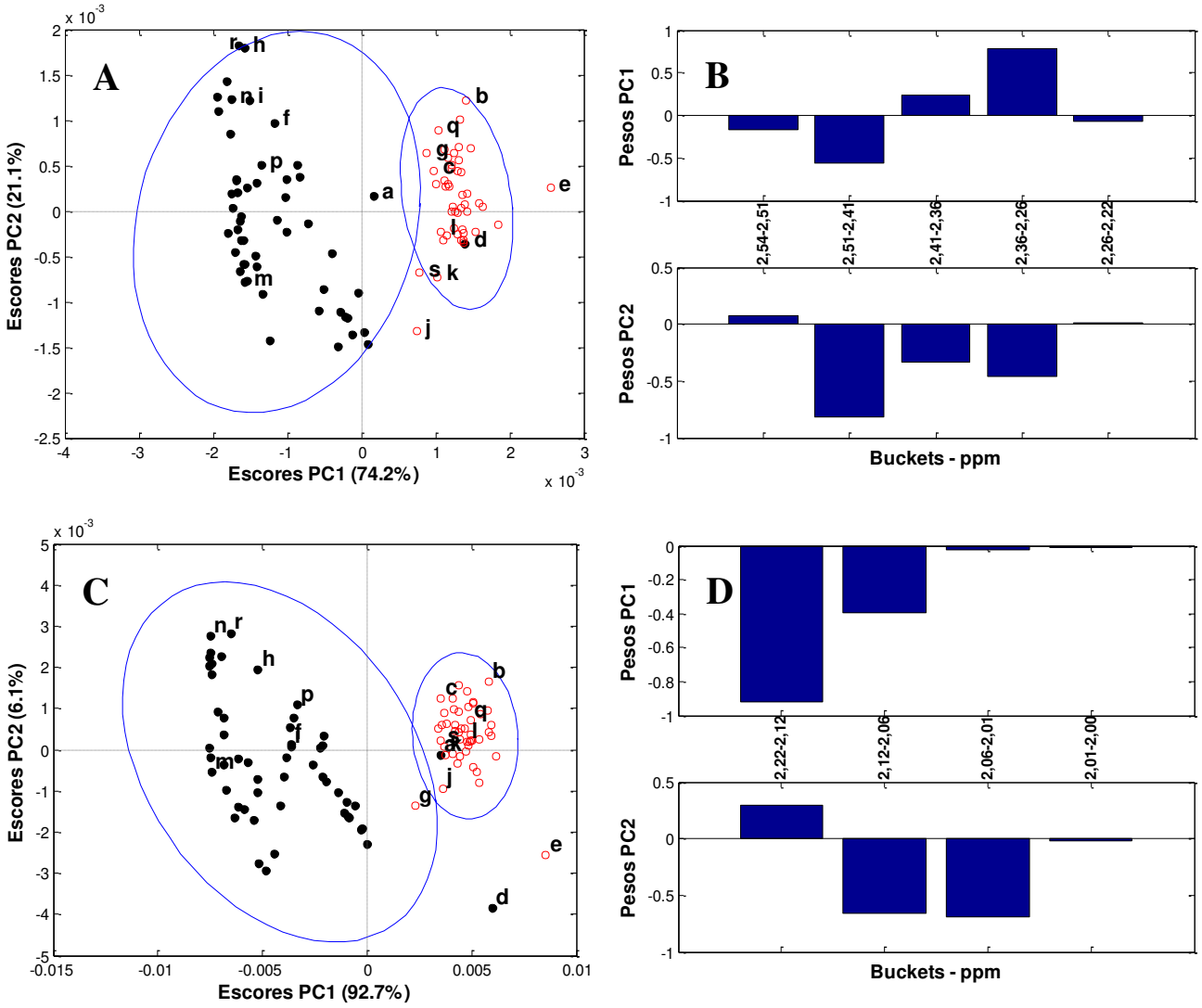

Figura 38: iPCA dos espectros de RMN ^{1}H. A e B, gráficos de escores e de pesos relativos ao intervalo 9; C e D, gráficos de escores e de pesos relativos ao intervalo 10. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

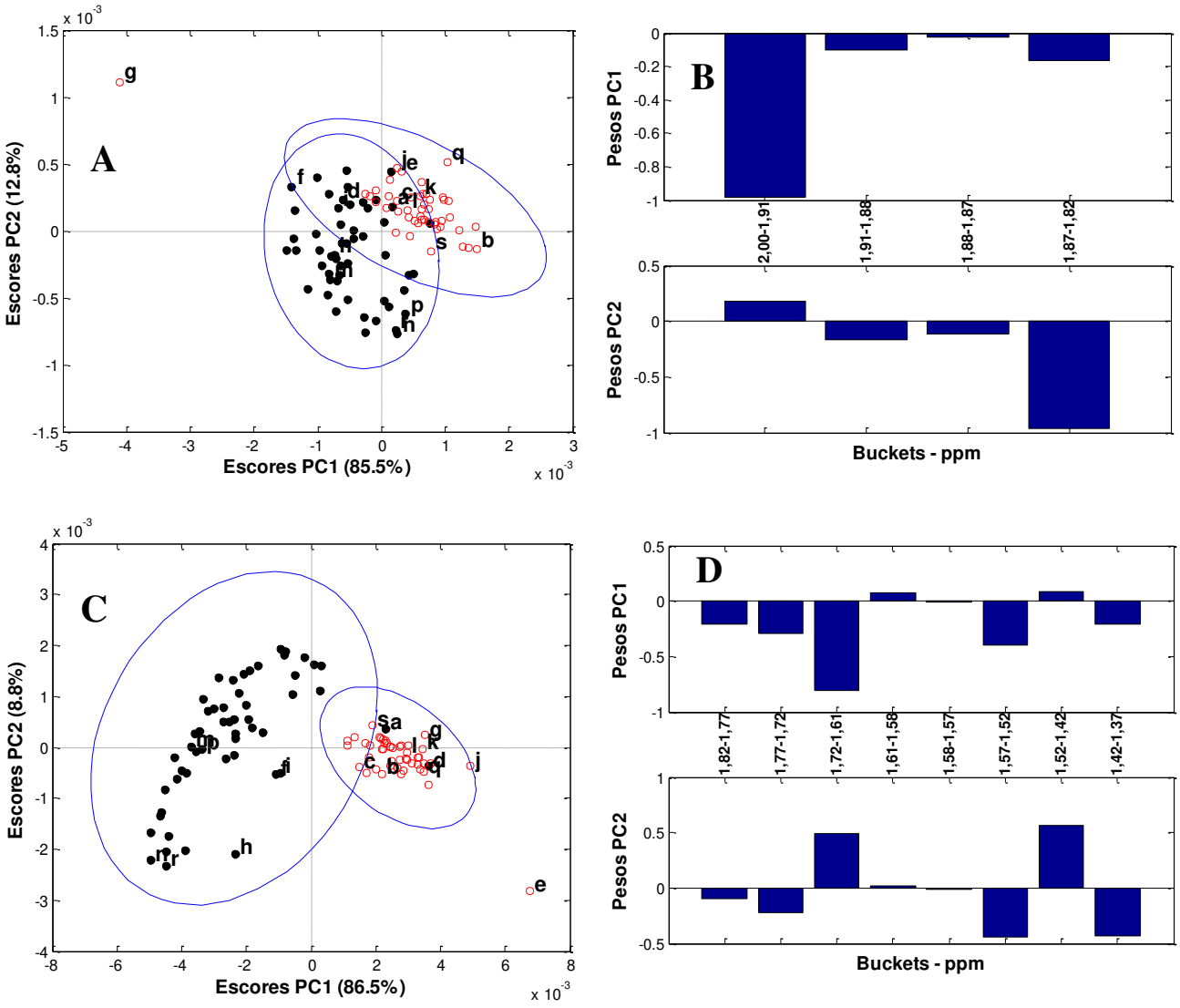

Figura 39: iPCA dos espectros de RMN ^{1}H. A e B, gráficos de escores e de pesos relativos ao intervalo 11; C e D, gráficos de escores e de pesos relativos ao intervalo 12. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

Como pode ser visto nas Figuras 38A e 38B (intervalo 9), pelos sinais em 2,51-2,41 ppm e nas Figuras 39C e 39D (intervalo 12), através dos sinais em 1,72-1,61 ppm, que se referem aos prótons nas posições α e β em relação aos grupos carbonílicos, respectivamente, e também nas Figuras 37C e 37D (intervalo 8), com sinais em 2,76-2,67 ppm e nas Figuras 39A e 39B, através dos sinais em 2,00-1,91 ppm, que são relativos aos hidrogênios nas posições bis-alílicas e alílicas,

respectivamente, nas cadeias insaturadas dos ésteres, é provável que as diferentes classes tenham diferentes fontes de biodiesel, uma vez que estes sinais estão envolvidos na discriminação em PC1 nestes intervalos (cada PC1 explica mais de 74% de variância em cada caso).

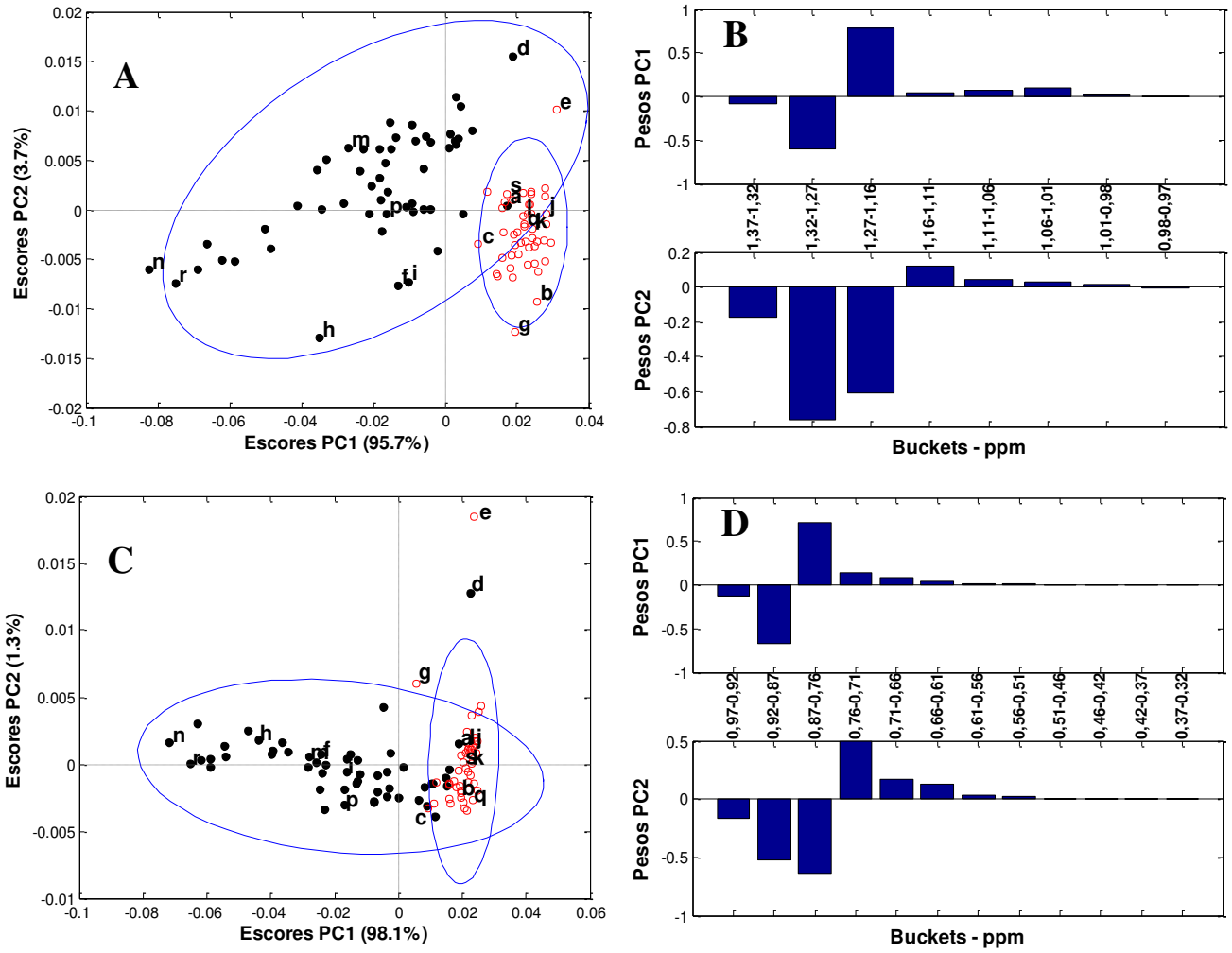

Figura 40: iPCA dos espectros de RMN ^{1}H. A e B, gráficos de escores e de pesos relativos ao intervalo 13; C e D, gráficos de escores e de pesos relativos ao intervalo 14. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

Os deslocamentos químicos em 1,32-1,27 ppm (Figuras 40A e 40B, intervalo 13) dos hidrogênios metilênicos das *iso*-parafinas (hidrocarbonetos saturados com

cadeias ramificadas) com pesos negativos em PC1 (95,7% de variância explicada) e os deslocamentos químicos em 1,27-1,16 ppm (Figuras 40A e 40B, intervalo 13) dos hidrogênios metilênicos das *n*-parafinas (hidrocarbonetos saturados com cadeias normais) com pesos positivos em PC1, indicam que as amostras metropolitanas, com escores negativos, têm mais alto conteúdo de *iso*-parafinas e as amostras não metropolitanas, com escores positivos, têm mais alto conteúdo de *n*-parafinas. Adicionalmente, os sinais em 1,57-1,52 ppm (Figuras 39C e 39D, intervalo 12) que representam os prótons dos grupos –CH– das *iso*-parafinas corroboram o maior teor destas substâncias para as amostras metropolitanas. O intervalo 12 (Figuras 39C e 39D) informa ainda sobre os naftenos (hidrocarbonetos cíclicos e saturados), sendo as amostras metropolitanas com maior conteúdo destes de acordo com a mesma análise citada para a PC1.

De modo geral, observa-se em quase todos os intervalos analisados nos espectros de RMN ^{1}H que as amostras não metropolitanas são mais homogêneas, dados os agrupamentos com menores desvios nos valores dos escores, ou seja, uma menor variância, o que inclusive resulta em elipses com 95% de confiança com eixos menores. Esta característica pode ser resultado da adição dos corantes vermelhos que para alguns intervalos "tornam" os espectros das misturas não metropolitanas mais similares. No entanto, dada a maneira com que os espectros foram adquiridos (baixo número de varreduras), é possível que a técnica não seja sensível à concentração de corante adicionado. Portanto, é mais provável que o tratamento para a redução do teor de enxofre nas amostras metropolitanas resulte numa maior variedade de espécies presentes nestas amostras, tornando-as assim mais heterogêneas. Aliado a isto, a adição de produtos derivados do *cracking* do petróleo a estas últimas, que visem à melhoria das propriedades combustíveis, podem favorecer as diferenças.

Com o objetivo de verificar a concordância da análise química acima com os parâmetros mostrados na Tabela 3, um modelo PCA foi construído para a matriz

autoescalada relativa a esta tabela, sendo mostrado na Figura 41. Pode ser visto que apenas a primeira componente principal evidencia diferenças entre as duas classes de amostras, apontando as respostas das propriedades: índice de cetano (IC); temperaturas de destilação de 50% (T50%) e 85% (T85%) do volume da amostra; massa específica (ME); teor de enxofre (Enxofre); e em menor extensão ponto de fulgor (PtFulg) como as responsáveis por esta discriminação, sendo as amostras não metropolitanas aquelas com os maiores valores destes parâmetros (em PC1, as amostras não metropolitanas possuem escores positivos e os parâmetros citados possuem pesos positivos).

Este resultado está de acordo com os mostrados pelo modelo iPCA acima, pois conforme constatado pelas análises dos intervalos, as amostras não metropolitanas possuem um maior conteúdo de aromáticos polinucleares, que por serem mais pesados e menos voláteis favorecem uma maior massa específica, maiores T50% e T85% e maior IC. De fato, as amostras metropolitanas tendem a possuir menores T50% e T85%, pois apresentam mais *iso*-parafinas e naftenos que são mais voláteis do que os respectivos hidrocarbonetos de cadeia normal. Adicionalmente, uma vez que *iso*-parafinas possuem temperaturas de autoignição maiores, elas favorecem um menor IC, parâmetro associado à qualidade de ignição no motor à diesel, o que resulta nos menores ICs das amostras metropolitanas frente às não metropolitanas, cujo conteúdo maior de *n*-parafinas (com menores temperaturas de autoignição) reforçam o resultado.

O ponto de fulgor está relacionado à inflamabilidade (o seu valor é levado em conta no manuseio, estocagem e transporte do combustível) tendo uma relação direta com a T50%, ambos diminuindo com a presença de substâncias mais voláteis (frações de hidrocarbonetos mais leves). Isto explica a relação desta variável na PCA, ao longo da PC1.

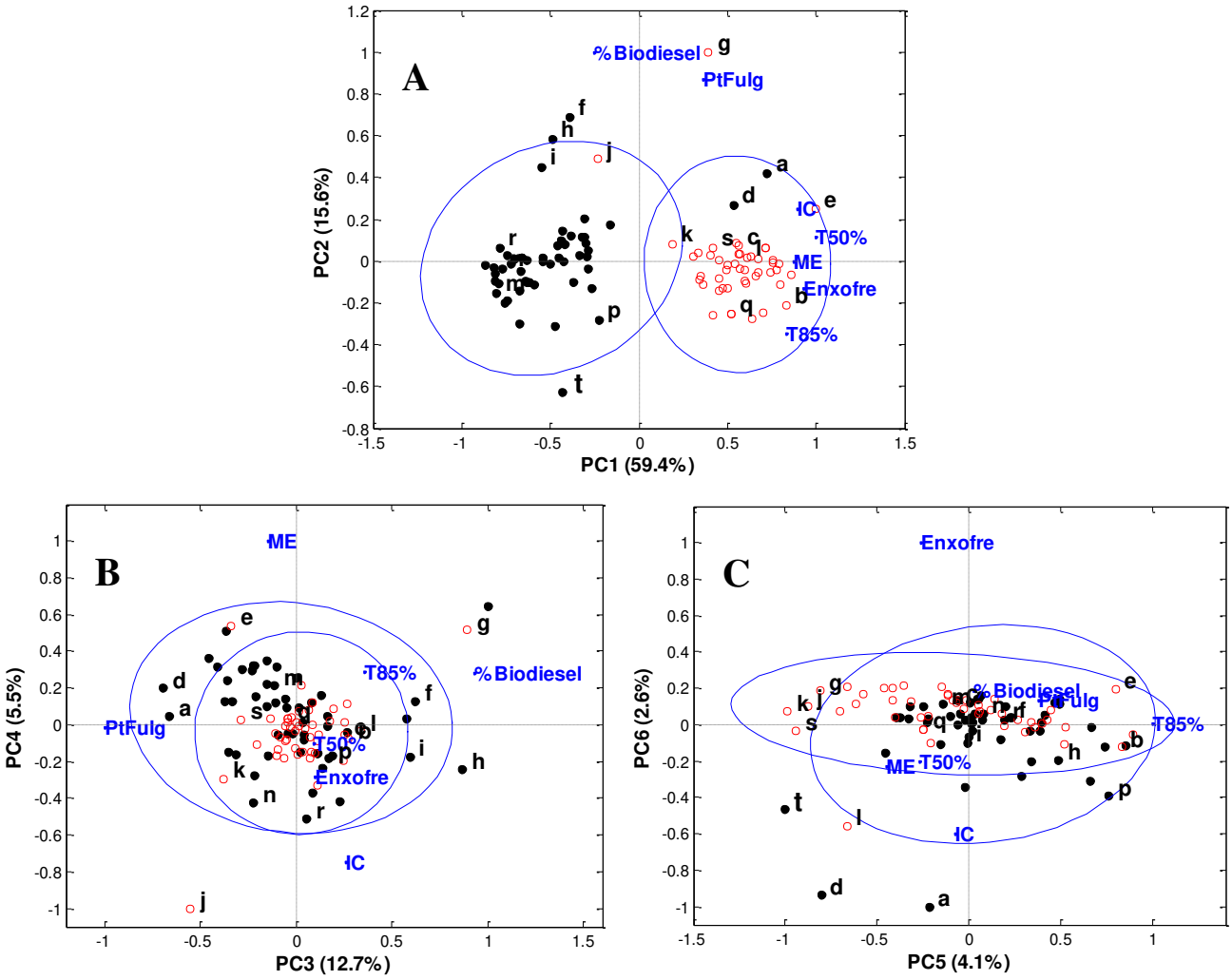

Figura 41: PCA para os parâmetros da Tabela 3. A: PC1 *versus* PC2; B: PC3 *versus* PC4; C: PC5 *versus* PC6. As amostras são identificadas do mesmo modo que na iPCA para os espectros de RMN ^{1}H. Variáveis: teor de biodiesel (%Biodiesel); índice de cetano (IC); temperaturas de destilação de 50% (T50%) e 85% (T85%) do volume da amostra; massa específica (ME); teor de enxofre (Enxofre); ponto de fulgor (PtFulg). Os gráficos são mostrados como *biplots* onde os escores e os pesos são normalizados com o objetivo de demonstrar a relação entre as amostras e variáveis diretamente.

As características químicas das misturas podem ainda ser correlacionadas a outras propriedades como: ponto de nuvem (PN - temperatura em que ocorre a cristalização das parafinas do combustível), ponto de entupimento de filtro a frio (PEFF - temperatura em que o combustível entope em um dispositivo de filtro padrão), viscosidade, estabilidade à oxidação, entre outras. Dentre estas, a

viscosidade, o PN e o PEFF são fortemente dependentes das características do biodiesel, especialmente do grau de instauração. As cadeias monoinsaturadas aumentam a viscosidade, enquanto as cadeias poli-insaturadas a diminuem. Além disso, um maior grau de instauração diminui o PN, o PEFF e a estabilidade à oxidação [104,105], o que não é desejável. O PN e o PEFF também dependem dos hidrocarbonetos presentes no combustível, sendo favorecidos os maiores valores, pela presença de hidrocarbonetos mais ramificados [103]. Infelizmente, os valores destes parâmetros não estão disponíveis para as amostras estudadas aqui, não permitindo suas correlações com a iPCA dos espectros de RMN ^{1}H. Apesar disso, pode-se prever que as amostras metropolitanas são menos estáveis à oxidação devido ao maior teor de insaturados. Para estas mesmas amostras, o maior grau de instauração e maior conteúdo de hidrocarbonetos ramificados (*iso*-parafinas) atuam opostamente para o PN e o PEFF dificultando alguma previsão.

No que diz respeito às características químicas das amostras não conformes, a análise por iPCA dos espectros de RMN ^{1}H permite algumas inferências, tendo em vista as interpretações dos pesos em cada intervalo, observando em qual intervalo a amostra foi detectada e o valor do seu escore. Desse modo, a presença de adulterantes nas misturas pode ser especulada e servir como um guia para investigações posteriores. Por exemplo: espera-se para a amostra fora da especificação “a” (metropolitana) um maior conteúdo de compostos aromáticos polinucleares (intervalos 1 e 2), biodiesel com características similares àqueles usados nas amostras não metropolitanas (intervalos 8, 9 e 12), e mais *n*-parafinas do que *iso*-parafinas (intervalo 13), tudo isto resultando numa maior densidade para esta amostra, além de maiores temperaturas de destilação (especialmente T50%), maior ponto de fulgor e índice de cetano (veja Tabela 3); a amostra metropolitana reprovada “d” é bastante diferente das outras da sua classe, com teor de biodiesel muito baixo (intervalo 6) e altos níveis de *n*-parafinas (intervalos 13 e 14), tendo assim alta densidade (os

hidrocarbonetos de cadeias normais tendem a ser mais densos do que os respectivos de cadeia ramificada). Essa amostra é ainda descrita praticamente sozinha pela resposta dos sinais em 3,96-3,85 ppm, ao longo de PC1 e PC2 no intervalo 5 (Figuras 36A e 36B); a amostra não metropolitana "g" possui o mais alto teor de biodiesel (intervalo 6), talvez por isso, o mais alto grau de insaturações (intervalo 4) e uma bastante distinta composição de hidrocarbonetos (intervalos 11, 13 e 14).

Para as amostras aqui investigadas os modelos locais favorecem uma melhor interpretação, pois a composição química peculiar estudada apresenta distintos tipos de compostos (aromáticos, alifáticos, cadeia aberta, cadeia fechada, cadeia normal, cadeia ramificada, entre outros) que mostram sinais de RMN em regiões específicas.

3.5.3. Análise exploratória por iPCA: espectros de RMN ^{13}C

Antes da análise exploratória dos espectros de RMN ^{13}C, os mesmos foram submetidos à redução de ruído utilizando a MSPCA, onde dois níveis de decomposição foram usados na transformada *wavelet*, quatro e cinco componentes principais para reconstruir os detalhes e seis componentes principais para a PCA final, conforme descrito nos itens 3.4.4.1 e 3.4.4.2. As Figuras 42A e 42B mostram os espectros em questão, na mesma escala, antes e depois do procedimento de redução no ruído, onde pode ser visto o bom desempenho do MSPCA para realização desta tarefa, pois através das regiões ampliadas 1, 2 e 3 (os números destas regiões não possuem relação com os números dos intervalos na iPCA) em cada uma das figuras, é visível a melhoria da relação sinal/ruído.

A relação sinal/ruído tende a ser importante apenas para a análise dos sinais menores, uma vez que o ruído pode mascará-los dado a proximidade das variâncias. Isto toma uma importância maior neste trabalho devido ao uso da análise de componentes principais em intervalos, pois ao se usar intervalos como aqueles nas

regiões ampliadas na Figura 42A, a variância dos sinais pode ser confundida com a variância do ruído e assim dificultar as inferências. Por outro lado, no uso dos espectros inteiros centrados na média e numa PCA convencional, os picos pequenos também tendem a ser mascarados, porém agora pelos outros sinais maiores.

Após a redução de ruído, os espectros de RMN ^{13}C foram divididos em um conjunto de treinamento e um de validação, sendo este último composto pelas amostras reprovadas pelos parâmetros na Tabela 3 e pelas amostras reprovadas pelos modelos iPCA para os espectros e hidrogênio. As misturas do conjunto de validação foram centradas na média das amostras de treinamento e projetados no subespaço gerado pelos pesos do modelo iPCA construído com estas últimas, obtendo-se a predição dos escores para as primeiras. As Figuras 43 e 44 mostram as elipses com 95% de confiança para os intervalos 2, 3, 4, 7, 8, 10 e 11, juntamente com os escores preditos para as amostras reprovadas. Os intervalos mostrados são aqueles que mostram a detecção de amostras que ainda não haviam sido reprovadas pelo modelo iPCA para os espectros de hidrogênio.

De modo geral, o modelo iPCA para os espectros de carbono mostra-se limitado em relação à detecção fornecida pelos modelos a partir dos espectros de hidrogênio, não detectando várias amostras (“i”, “j”, “k”, “m”, “q”, “s”, “t”, “u”, “v”, “aa”, “cc”, “ff”, “hh”, “jj”, “kk”, “mm” e “nn”). Por outro lado, o modelo apontou quatro “novas” amostras não conformes, não indicadas anteriormente. Tais amostras são: 18 (“pp”) nos intervalos 2 (Figura 43A) e 3 (Figura 43B); 63 (“qq”) nos intervalos 2 (Figura 43A), 4 (Figura 43C), 7 (Figura 43D), 8 (Figura 44A) e 11 (Figura 44C); 14 (“rr”) no intervalo 3 (Figura 43B); e 30 (“ss”) no intervalo 10 (Figura 44B).

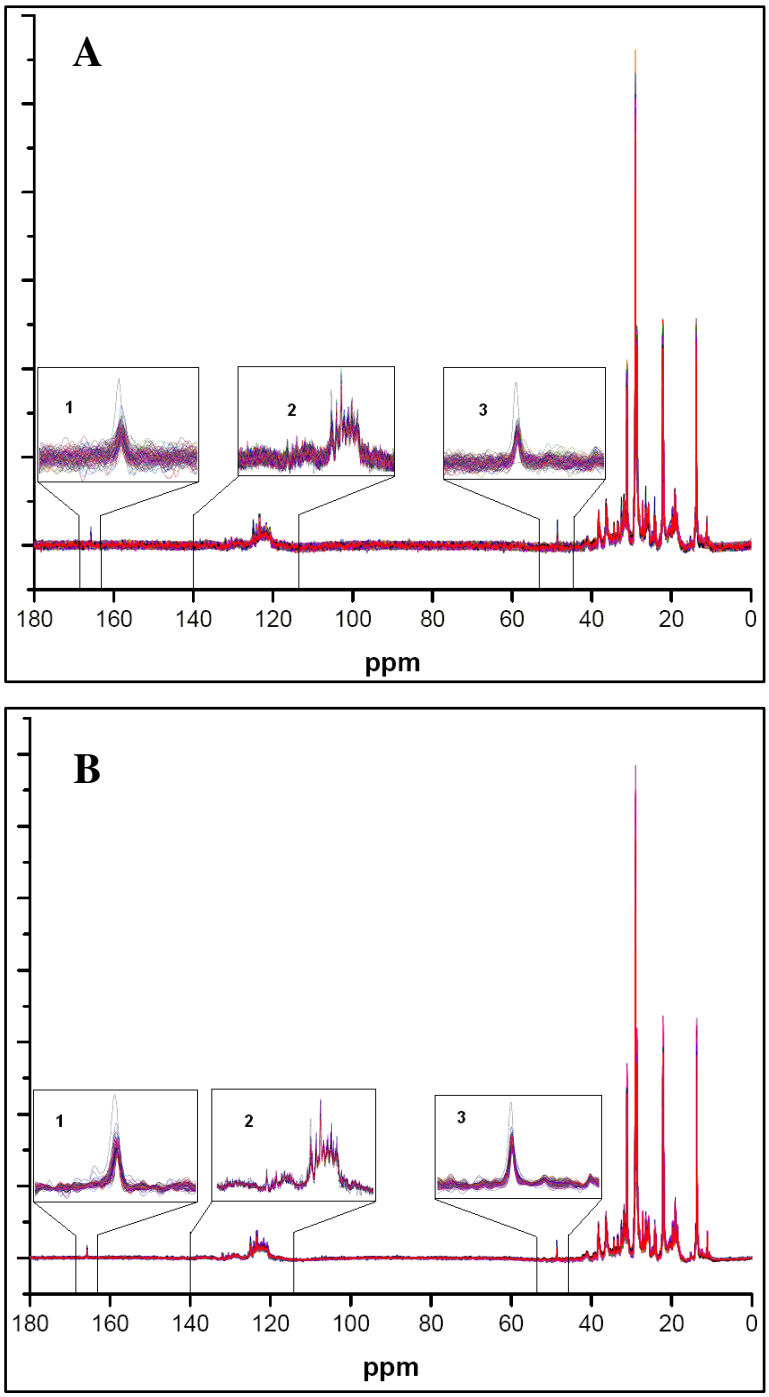

Figura 42: Espectros de RMN ^{13}C das 100 amostras de biodiesel-diesel, A: antes e B: depois da redução de ruído utilizando a abordagem de MSPCA, conforme descrito nos itens 3.4.4.1 e 3.4.4.2. Os espectros são mostrados na mesma escala nas ordenadas (intensidade de sinal).

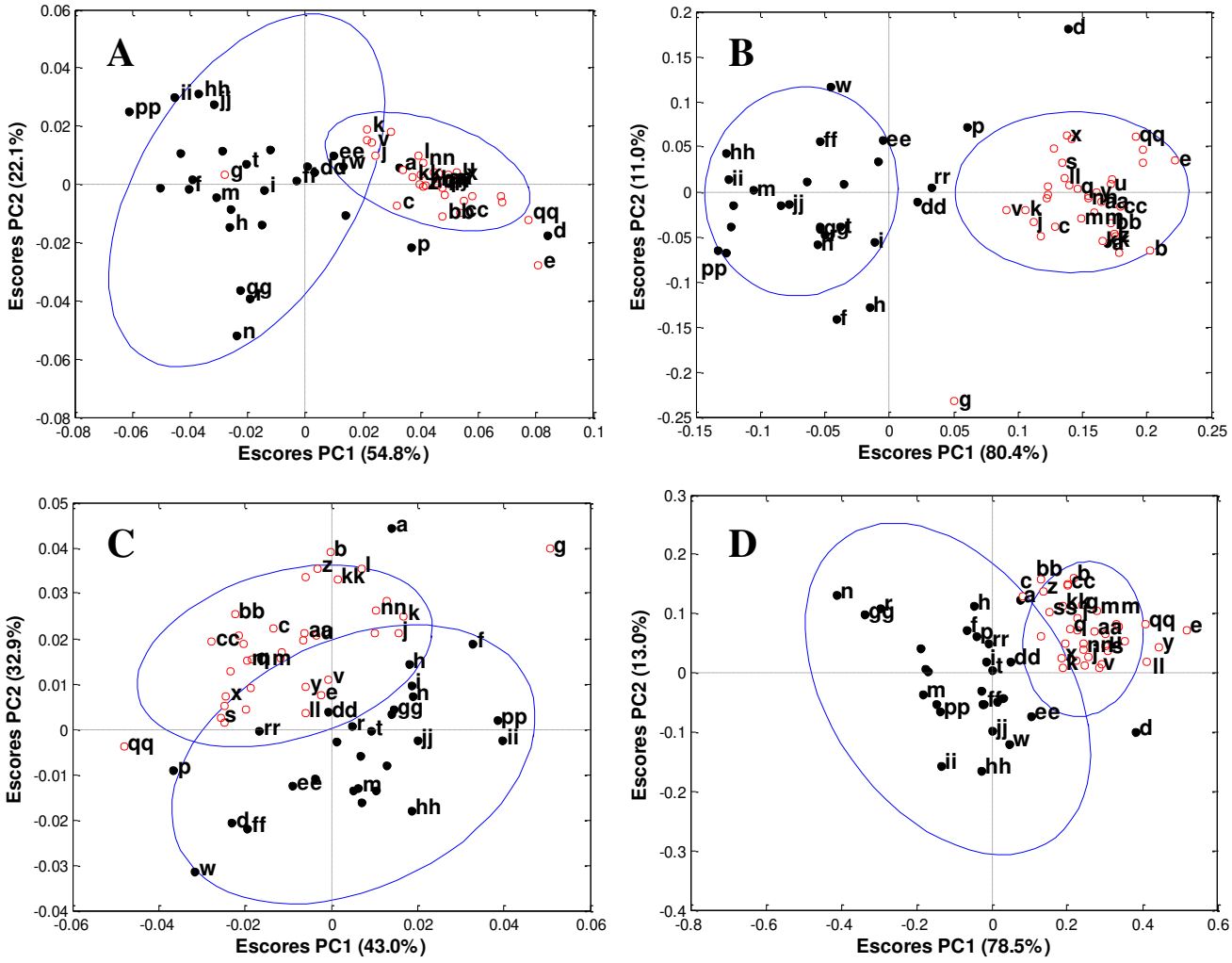

Figura 43: iPCA dos espectros de RMN ^{13}C das amostras não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A: gráfico de escores para o intervalo 2; B: gráfico de escores para o intervalo 3; C: gráfico de escores para o intervalo 4; D: gráfico de escores para o intervalo 7. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

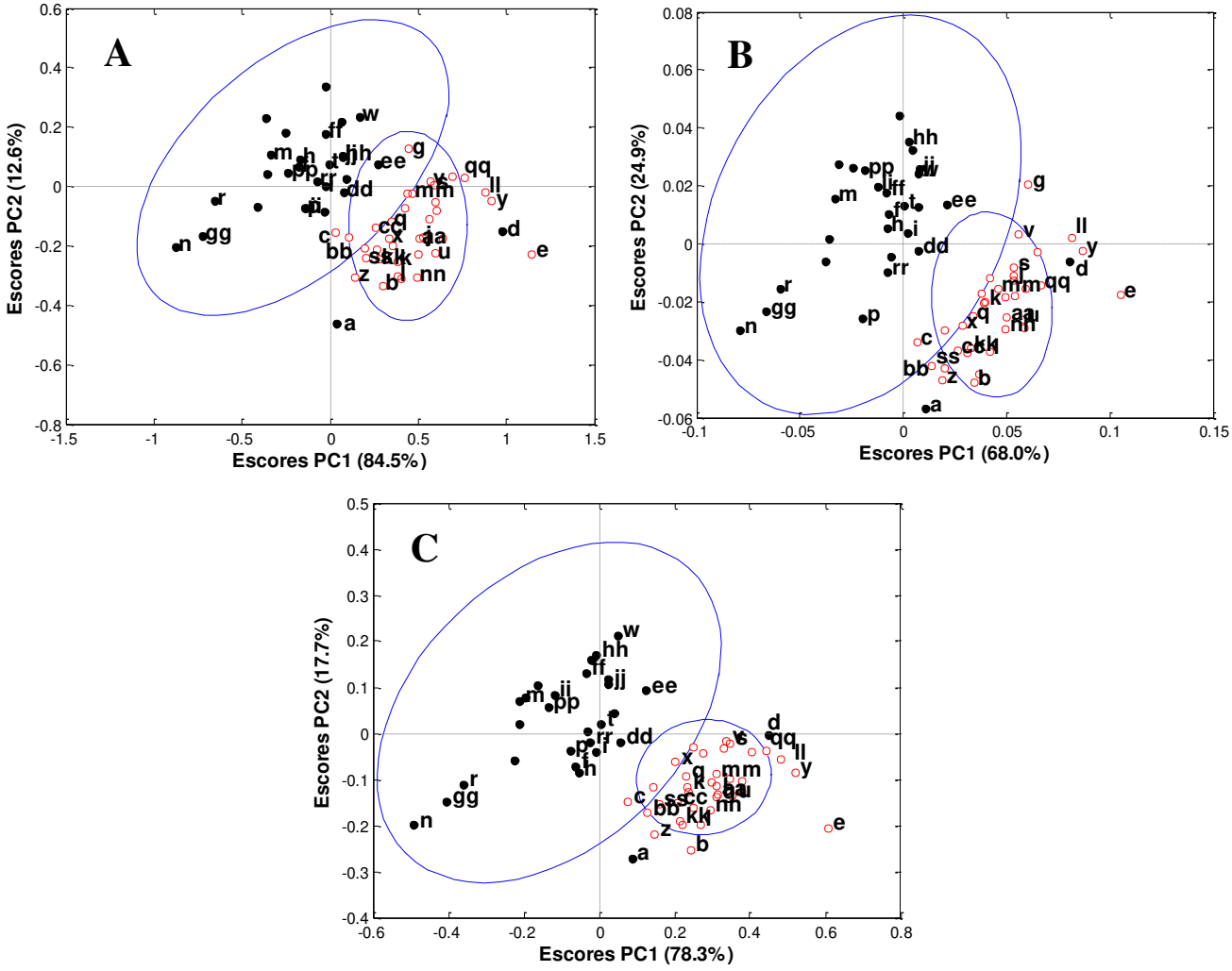

Figura 44: iPCA dos espectros de RMN ^{13}C das amostras não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A: gráfico de escores para o intervalo 8; B: gráfico de escores para o intervalo 10; C: gráfico de escores para o intervalo 11. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

As Figuras 45 a 52 mostram os modelos iPCA obtidos para cada classe de amostras separadamente, utilizando as amostras do mesmo conjunto de treinamento usado no modelo iPCA anterior. Novamente, as amostras do conjunto de validação são centradas na média das amostras de treinamento e assim projetadas nos subespaços gerados pelos pesos dos modelos.

O modelo evidencia uma "nova" amostra metropolitana de número 32 ("tt") nos intervalos 1 (Figura 45A), 3 (Figura 46A) e 4 (Figura 46C), sendo esta uma amostra não reprovada pelos parâmetros físico-químicos. Apesar disso, o modelo apresentou um pior desempenho quando comparado aos modelos com as duas classes juntas, onde algumas amostras antes detectadas, sendo algumas reconhecidamente fora da especificação, não mais foram detectadas ("l", "n", "r", "x", "dd", "ee" e "gg"). Isto significa, provavelmente, que a inclusão das variabilidades das duas classes é necessária para evidenciar as diferenças destas amostras, ou seja, dada a inclusão de amostras bastante distintas, o subespaço gerado pelo modelo tende a representar uma variação mais abrangente, capturando as variáveis que distinguem estas amostras.

Diante do exposto, reforça-se a sugestão deste trabalho, no que diz respeito à exploração de modelos iPCA com as duas classes juntas e modelos com estas separadas.

Por fim, pode ser mencionado que comparando os pré-tratamentos dos espectros de RMN aqui estudados, percebe-se que aquele utilizado para os espectros de RMN ^{13}C (diminuição do ruído) apresenta-se como uma tarefa mais complexa do que aquele realizado nos espectros de RMN ^{1}H (algoritmo de *bucketing* otimizado), possuindo um maior número de passos com necessidade de otimização (por exemplo, número de níveis de decomposição na DTW, critério de escolha, tipos de funções de base *wavelet*, número de componentes principais na MSPCA, etc.), o que pode ter influenciado no desempenho dos modelos.

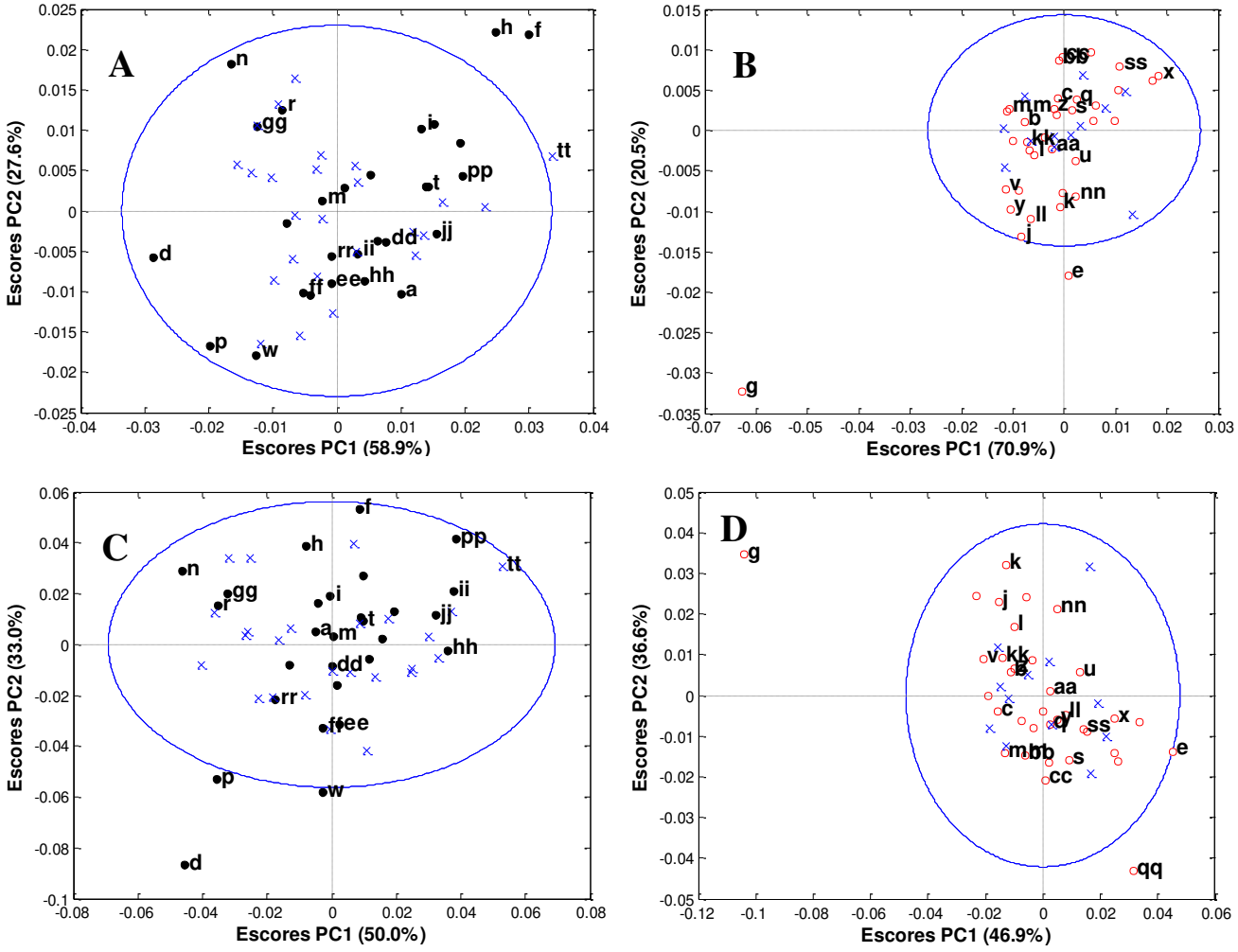

Figura 45: iPCA dos espectros de RMN ^{13}C para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 1) e C (intervalo 2): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 1) e D (intervalo 2): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

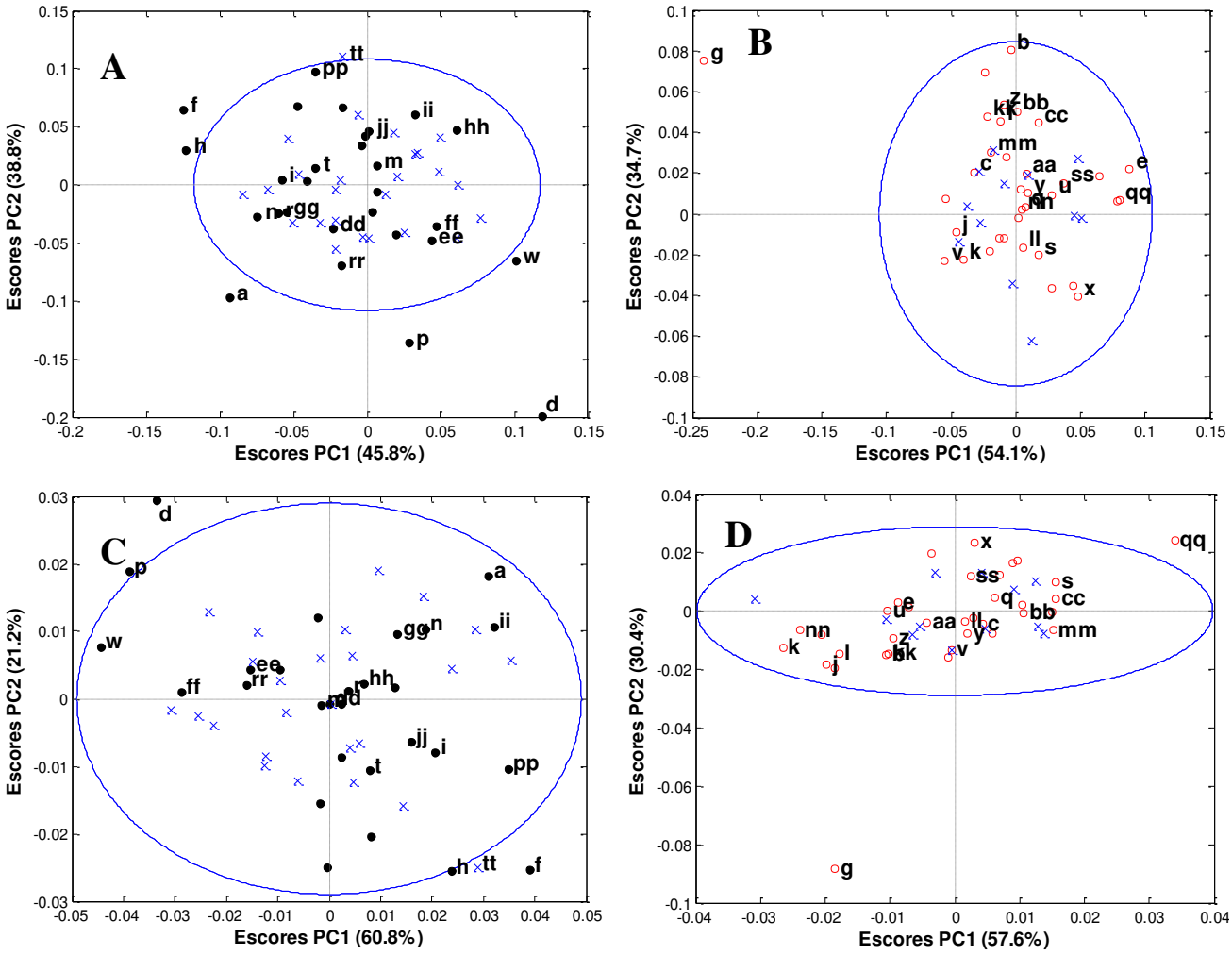

Figura 46: iPCA dos espectros de RMN ^{13}C para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 3) e C (intervalo 4): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 3) e D (intervalo 4): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

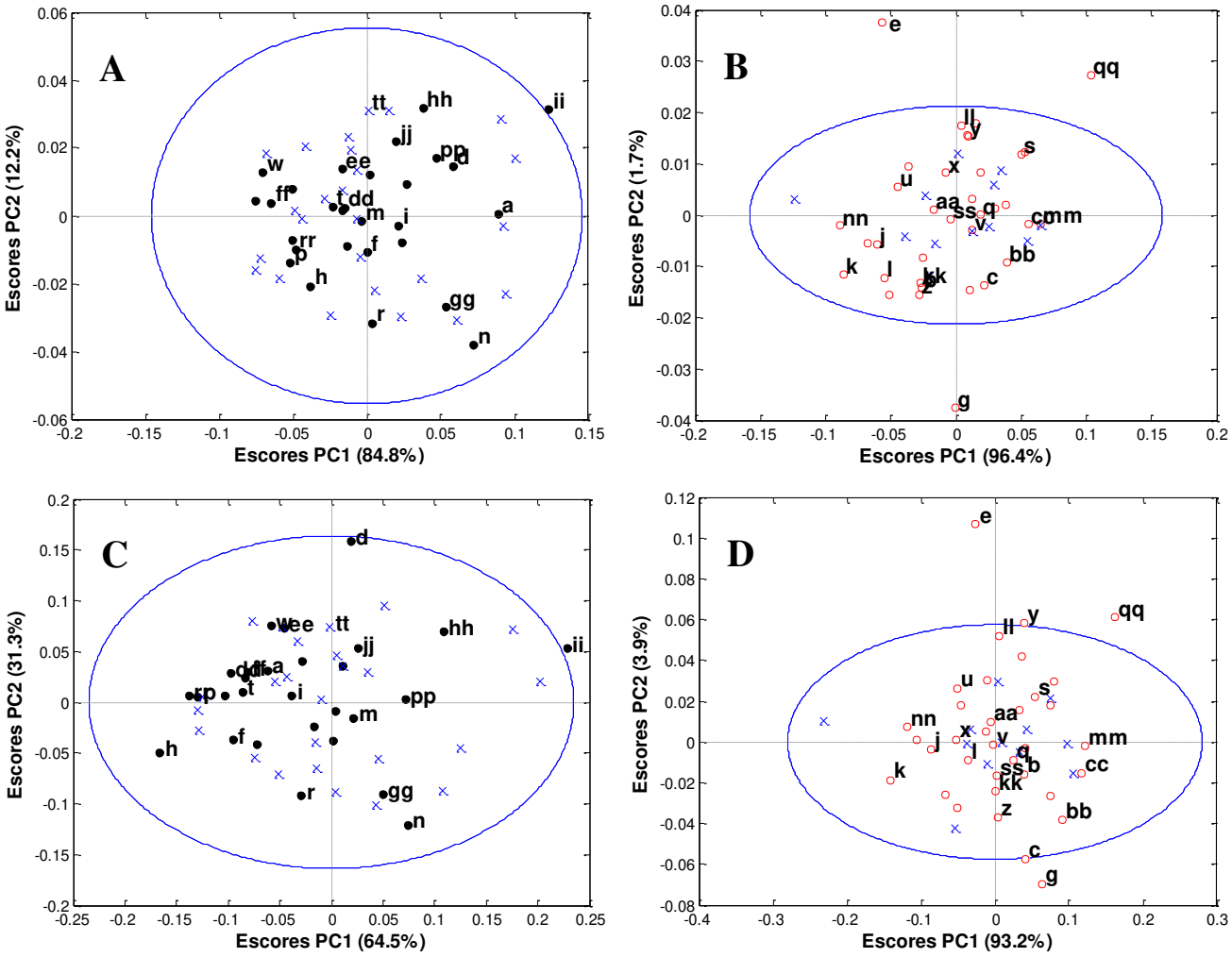

Figura 47: iPCA dos espectros de RMN ^{13}C para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 5) e C (intervalo 6): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 5) e D (intervalo 6): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

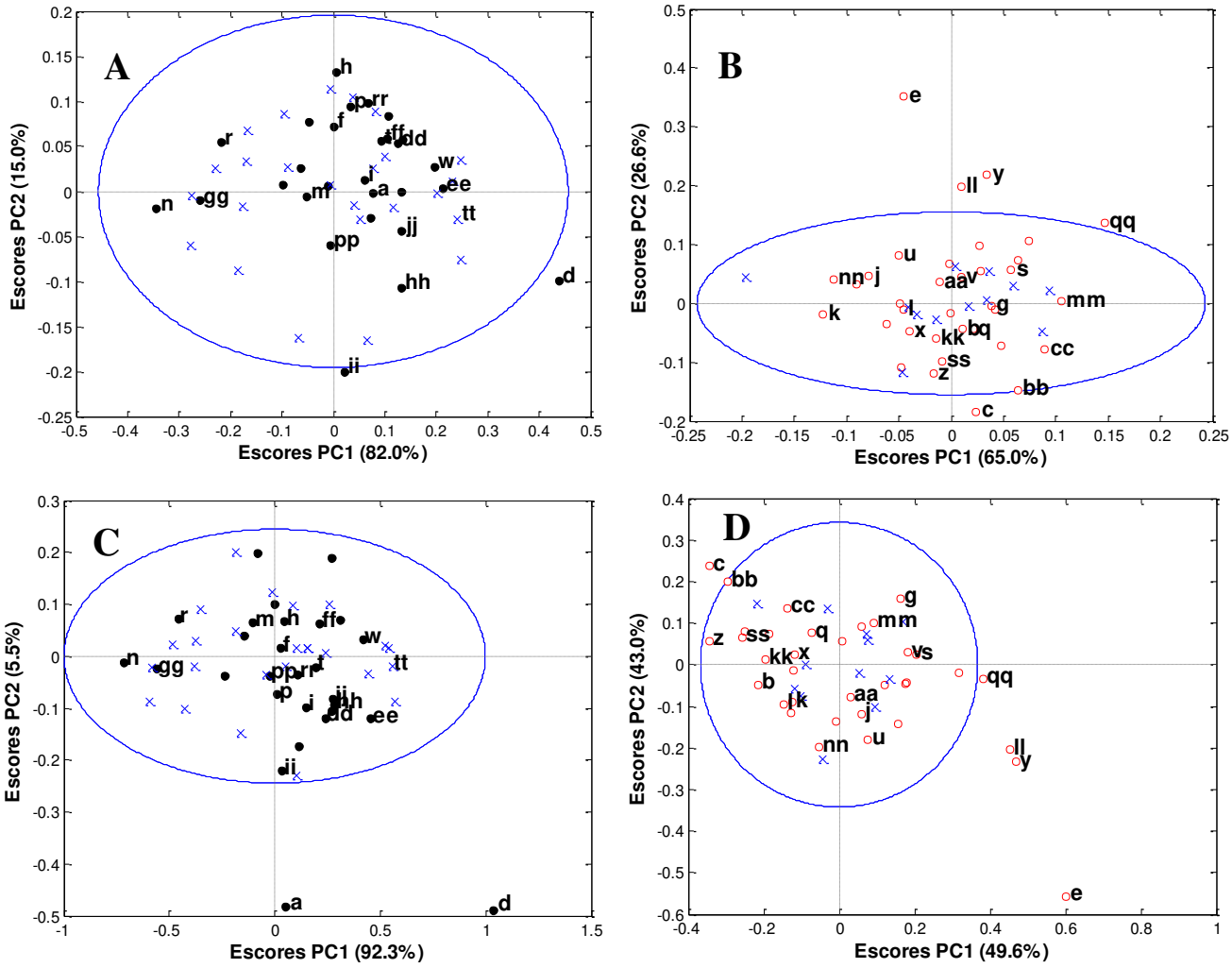

Figura 48: iPCA dos espectros de RMN ^{13}C para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 7) e C (intervalo 8): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 7) e D (intervalo 8): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

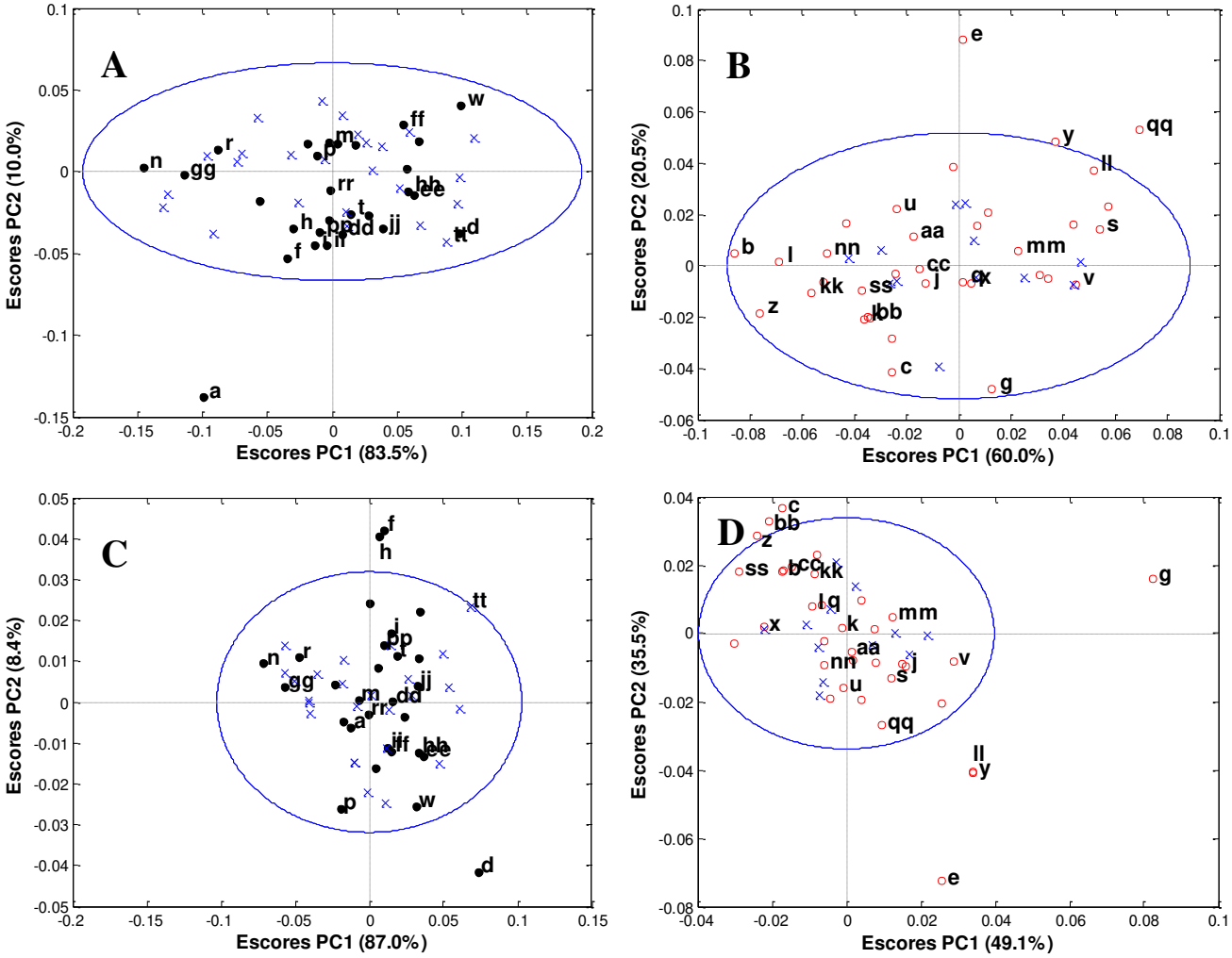

Figura 49: iPCA dos espectros de RMN ^{13}C para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 9) e C (intervalo 10): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 9) e D (intervalo 10): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

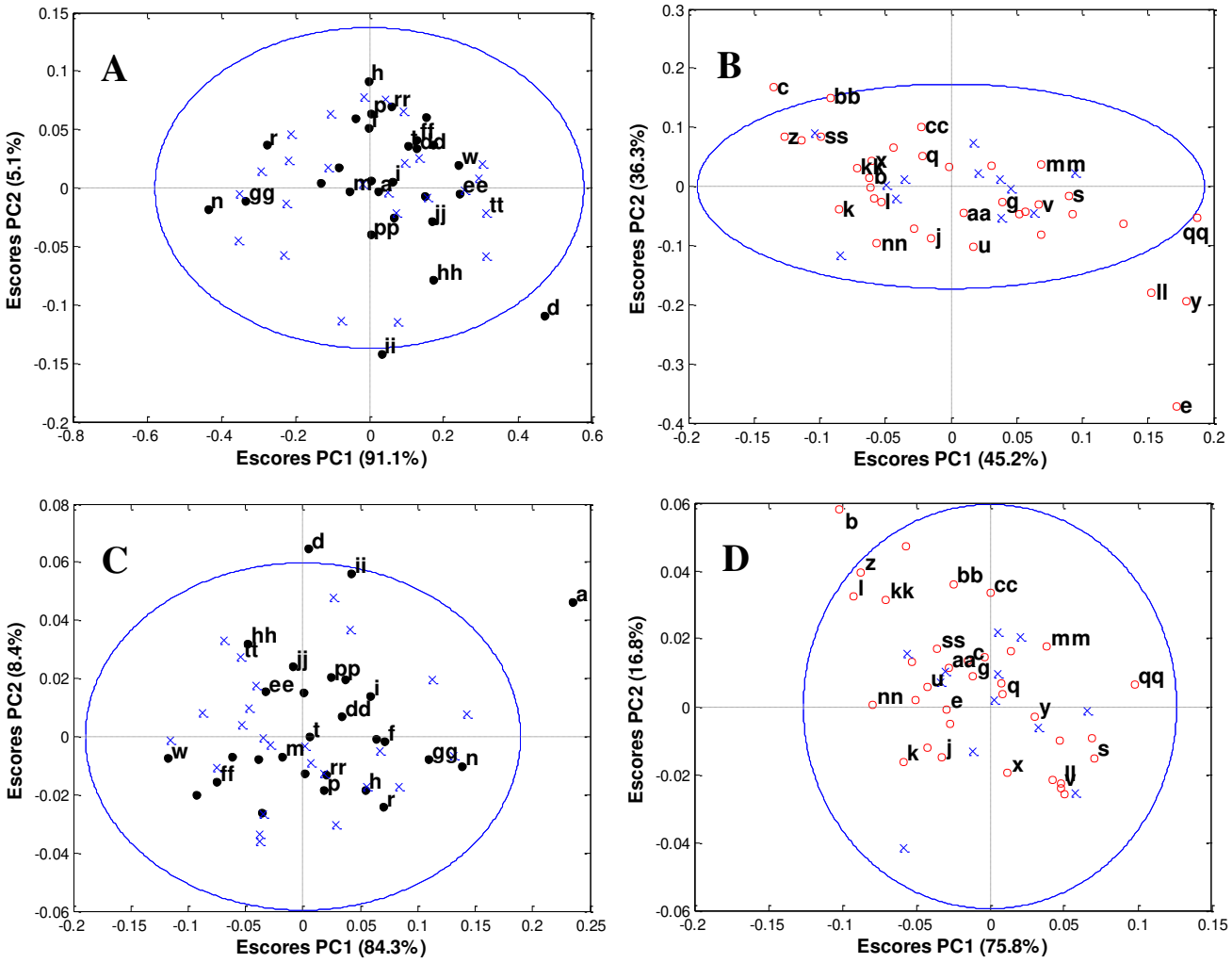

Figura 50: iPCA dos espectros de RMN ^{13}C para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 11) e C (intervalo 12): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 11) e D (intervalo 12): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

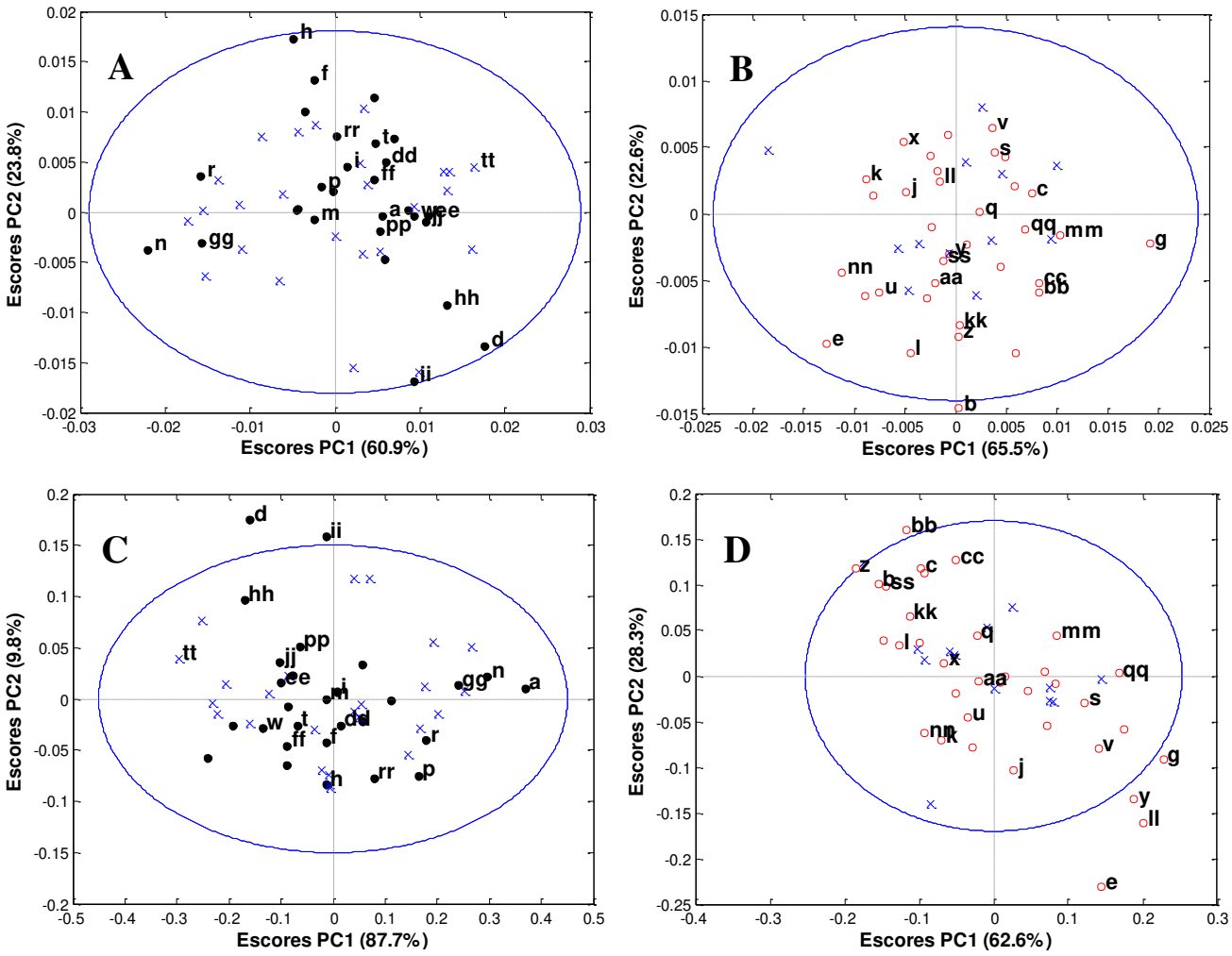

Figura 51: iPCA dos espectros de RMN ^{13}C para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 13) e C (intervalo 14): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 13) e D (intervalo 14): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

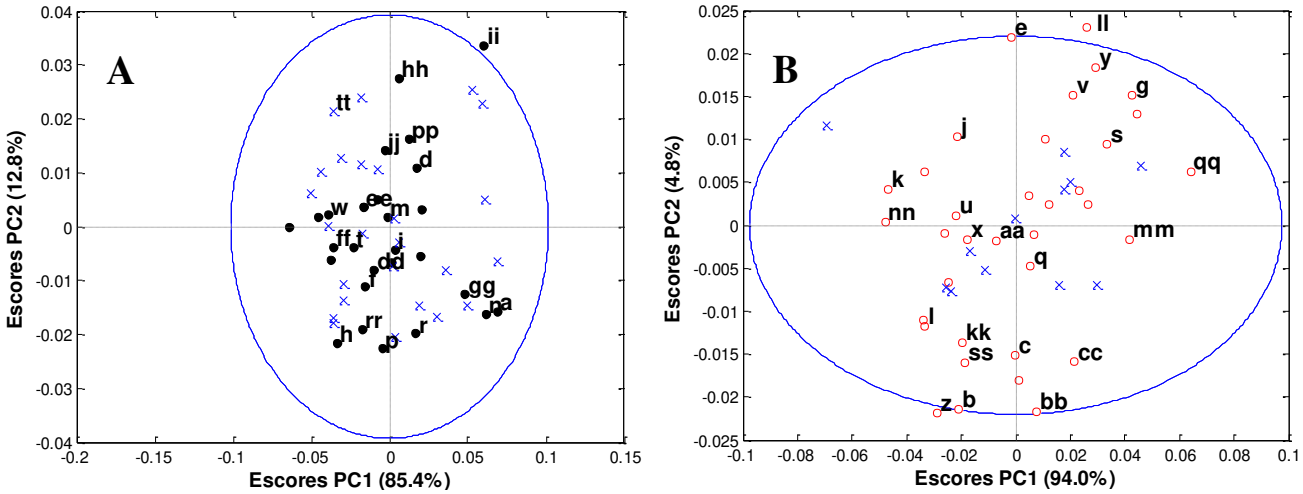

Figura 52: iPCA dos espectros de RMN ^{13}C para as classes de amostras de biodiesel-diesel não reprovadas. Elipses com 95% de confiança e escores preditos para as amostras reprovadas. A (intervalo 15): gráfico de escores para as amostras metropolitanas (pontos pretos); B (intervalo 15): gráfico de escores para as amostras não metropolitanas (círculos vermelhos). As cruzes azuis representam os escores das amostras não reprovadas utilizadas nos modelos.

3.5.3.1. Espectros de RMN ^{13}C – características químicas das classes

Os espectros de RMN ^{13}C, centrados na média e após a redução de ruído, das cem amostras de misturas de biodiesel-diesel foram utilizados para a construção de um modelo iPCA, sendo os resultados mostrados nas Figuras 53 a 60 onde as duas primeiras componentes principais são analisadas. Da mesma forma como ocorreu para os espectros de RMN ^{1}H, a iPCA mostrou separações entre as amostras metropolitanas e não metropolitanas, através da primeira componente principal (máxima variância) em quase todos os intervalos (em alguns intervalos essa diferença é mostrada na segunda componente principal) e também evidenciou diferentes misturas não conformes, ao longo dos diferentes intervalos analisados.

O modelo iPCA aqui estudado também possibilita inferências sobre as características químicas das classes de amostras que concordaram muito com a

mesma descrição fornecida pelo modelo iPCA dos espectros de RMN ^{1}H, inclusive indicando a maior homogeneidade das amostras na classe não metropolitana.

Nos gráficos de escores e pesos nas Figuras 53C e 53D e nas Figuras 54A e 54B (intervalos 2 e 3 para os espectros de RMN ^{13}C, respectivamente) pode ser notado que as amostras metropolitanas (escores negativos em PC1) têm maior teor de substâncias responsáveis pelos deslocamentos químicos entre 124 e 137 ppm (pesos negativos em PC1), que podem ser sinais de carbonos em compostos aromáticos ou insaturados (olefinas). Com relação ao teor de olefinas, este resultado concorda com a iPCA dos espectros de RMN ^{1}H nos intervalos: 4 (Figuras 35C e 35D), através dos sinais de prótons olefínicos; 8 (Figuras 37C e 37D), através dos sinais de hidrogênios bis-alílicos; 11 (Figuras 39A e 39B), através dos sinais de hidrogênios alílicos, sendo concordante também o destaque da amostra "g". As Figuras 55C e 55D (intervalo 6) "reafirmam" através dos pesos para os deslocamentos químicos entre 32,5 e 36 ppm que as amostras metropolitanas têm maior teor de olefinas, pois estes sinais correspondem a carbonos ligados a átomos de carbono com hibridização sp^2. Já a observação dos compostos aromáticos concorda com o intervalo 3 (Figuras 35A e 35B) da iPCA para os espectros de hidrogênio, através dos sinais dos prótons em compostos aromáticos mononucleares.

O intervalo 6 (Figuras 55C e 55D) para a iPCA dos espectros de carbono mostra, em PC2, a correlação das amostras não metropolitanas com os sinais em 37,8 e 39,7 ppm, relativos à ramificações de compostos aromáticos polinucleares e a correlação das amostras metropolitanas com os sinais em 39,5 e 39,9 ppm, associados às ramificações em compostos aromáticos mononucleares, exatamente como apontaram os espectros de RMN ^{1}H (intervalos 1 e 2 - Figura 34 e intervalo 3 - Figuras 35A e 35B).

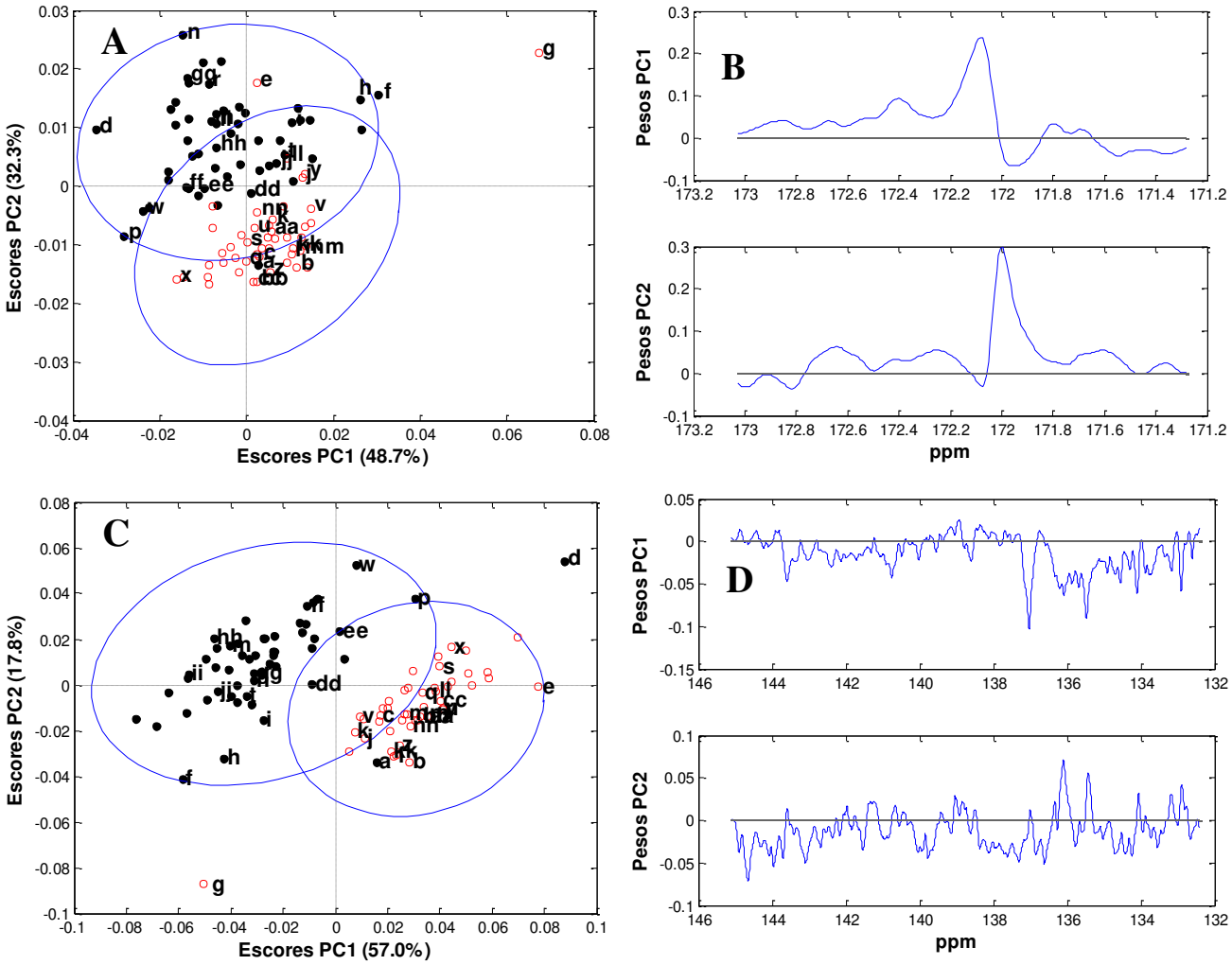

Figura 53: iPCA dos espectros de RMN ^{13}C. A e B, gráficos de escores e de pesos relativos ao intervalo 1; C e D, gráficos de escores e de pesos relativos ao intervalo 2. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

O intervalo 4 (Figuras 54C e 54D) mostra a relação das amostras com o sinal em 50,5 ppm, referente à metila ligada à porção dos ésteres que compõem os biodieseis, sendo, portanto, análogo ao intervalo 6 (Figuras 36C e 36D) na iPCA para os espectros de RMN ^{1}H. A amostra "g" destaca-se mais uma vez como aquela com maior teor de biodiesel, porém outras amostras com as mesmas características como as amostras "f" e "h" não são reprovadas na análise com os espectros de carbono. A PC2 no intervalo 4 indica possivelmente a diferença das classes, quanto à origem dos

biodieseis. Uma tendência similar, inclusive com relação à amostra "g", é observada através do intervalo 1 (Figuras 43A e 53B) que mostra o sinal em 172 ppm, referente aos grupos carbonílicos dos ésteres que compõem o biodiesel. Este intervalo também se mostra mais limitado na detecção de amostras com alto teor de biodiesel quando comparado ao intervalo 6 nas Figuras 36C e 36D. Adicionalmente, os sinais em 24,8; 25,3; 26,7; 27,2; 27,6; 28,3; 33,1 e 33,8 ppm, nos intervalos 9 e 10 (Figura 57) e no intervalo 6 (Figuras 55C e 55D) estão envolvidos na discriminação das classes ao longo da PC1 e uma vez que são atribuídos aos átomos de carbono nas cadeias dos ésteres, eles corroboram a inferência sobre as diferentes fontes de biodiesel cada classe.

Os deslocamentos químicos próximos de 23,0 (intervalo 11 – Figuras 58A e 58B); 29,8 e 30,2 (intervalo 8 - Figuras 56C e 56D); 32,3 (intervalo 7 - Figuras 56A e 56B) são característicos de carbonos (grupos metilênicos) nas cadeias longas de *n*-parafinas e *iso*-parafinas. A iPCA correlaciona os sinais à esquerda de cada um dos deslocamentos citados, que são relativos às *iso*-parafinas, com as amostras metropolitanas e os sinais à direita, que são relativos às *n*-parafinas, com as amostras não metropolitanas. Além disso, analisando-se os sinais entre 14 e 22 ppm (intervalo 12 - Figuras 58C e 58D; intervalos 13 e 14 – Figura 59) percebe-se uma maior variabilidade de ramificações para as amostras metropolitanas, uma vez que os picos se referem a metilas terminais. Estes resultados estão de acordo com as interpretações na iPCA dos espectros de hidrogênio (intervalo 12 - Figuras 39C e 39D; intervalo 13 - Figuras 40A e 40B).

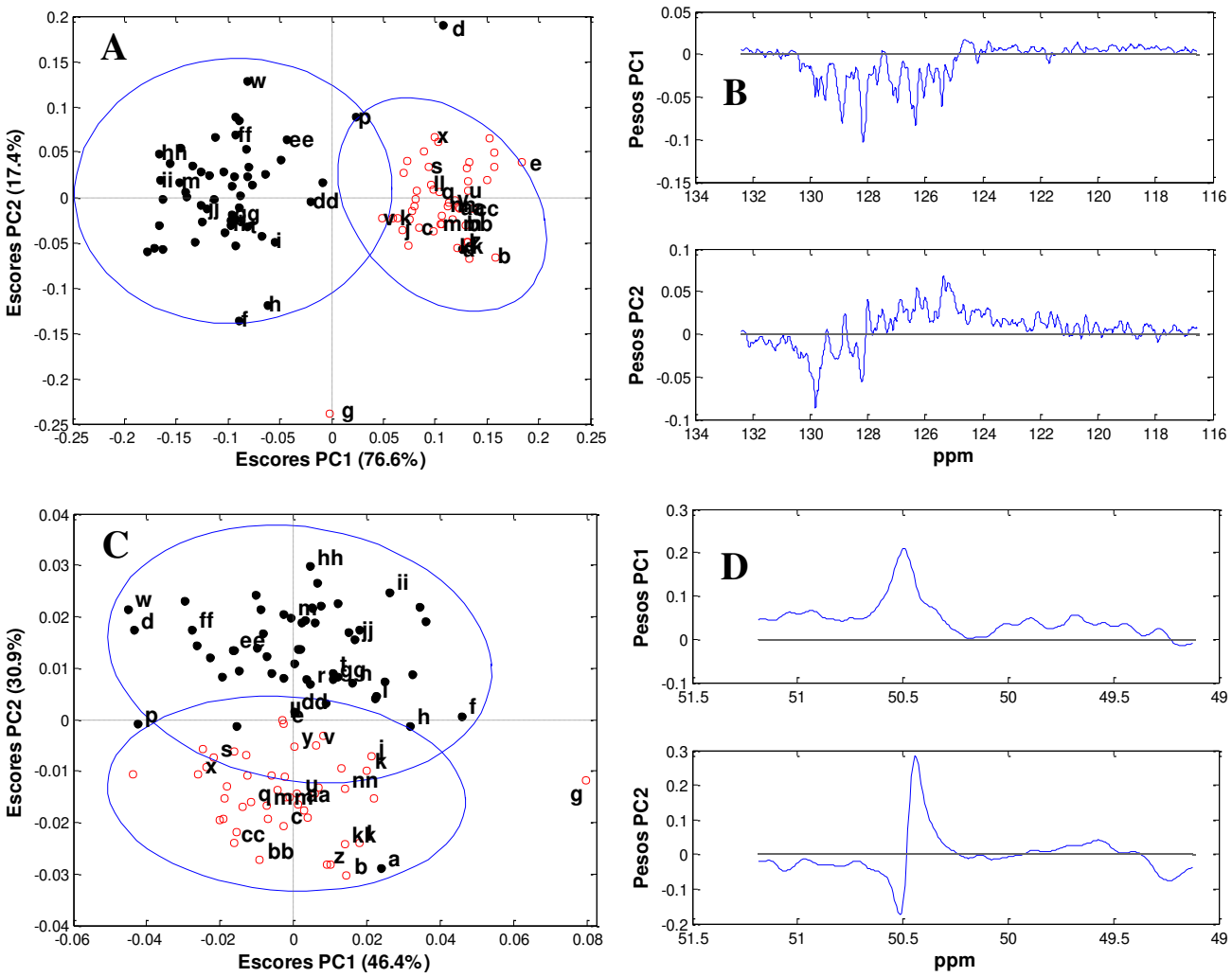

Figura 54: iPCA dos espectros de RMN ^{13}C. A e B, gráficos de escores e de pesos relativos ao intervalo 3; C e D, gráficos de escores e de pesos relativos ao intervalo 4. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

No que diz respeito à detecção das misturas fora da especificação, algumas amostras apontadas pelo modelo iPCA dos espectros de RMN ^{13}C também são indicadas pelo modelo a partir dos espectros de RMN ^{1}H e são rotuladas como: “a”, “b”, “d”, “e”, “f”, “g”, “l”, “n”, “p”, “y”, “z”, “ii” e “ll”. No entanto, com relação a todas as amostras identificadas, nas quais existem algumas reconhecidamente discrepantes, o modelo iPCA para os espectros de carbono mostrou-se mais limitado, não identificando, por exemplo, as amostras “h” e “i” com teor de biodiesel bem

maior que as demais e a amostra “t” com ponto de fulgor menor dentre todas (veja Tabela 3).

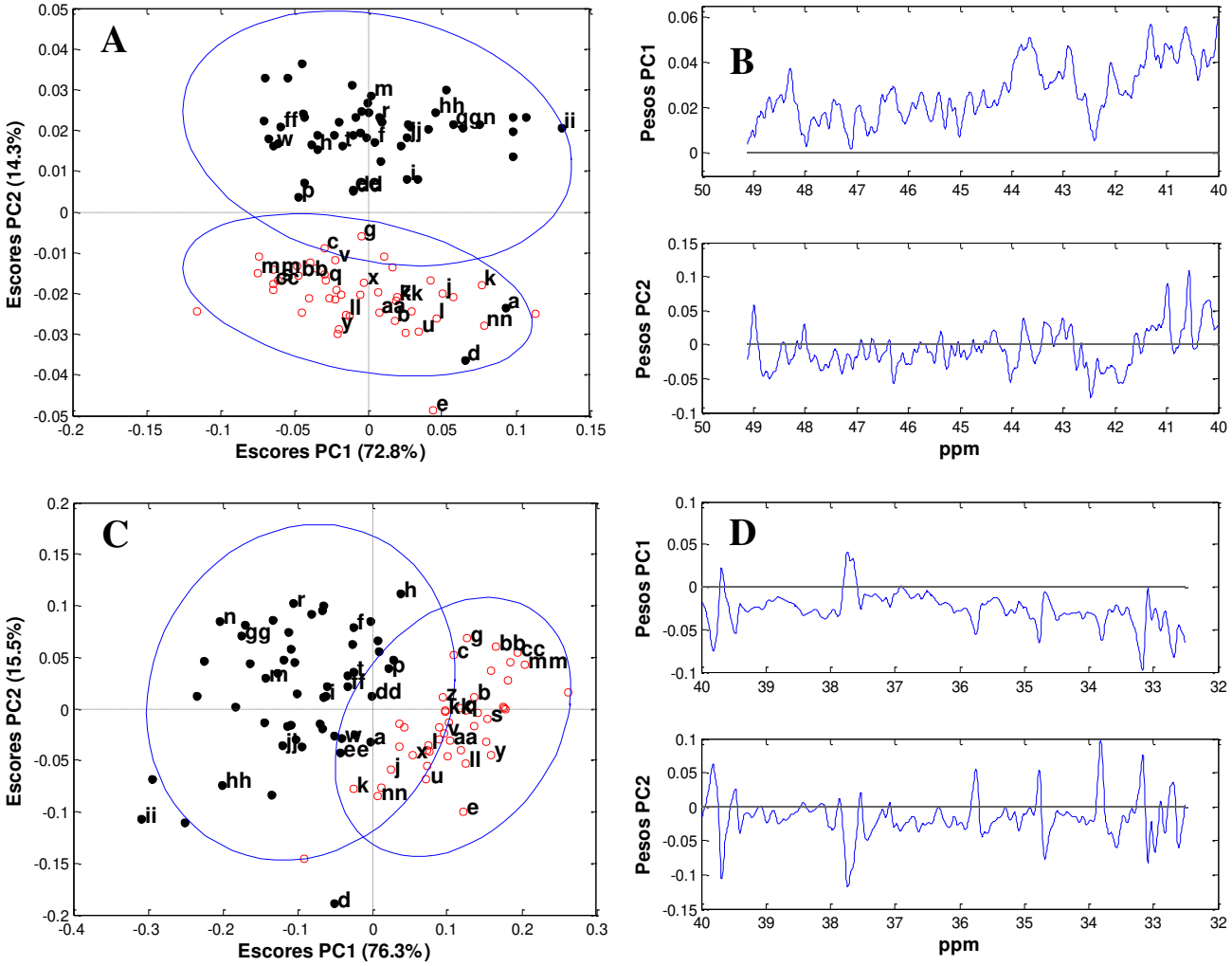

Figura 55: iPCA dos espectros de RMN ^{13}C. A e B, gráficos de escores e de pesos relativos ao intervalo 5; C e D, gráficos de escores e de pesos relativos ao intervalo 6. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

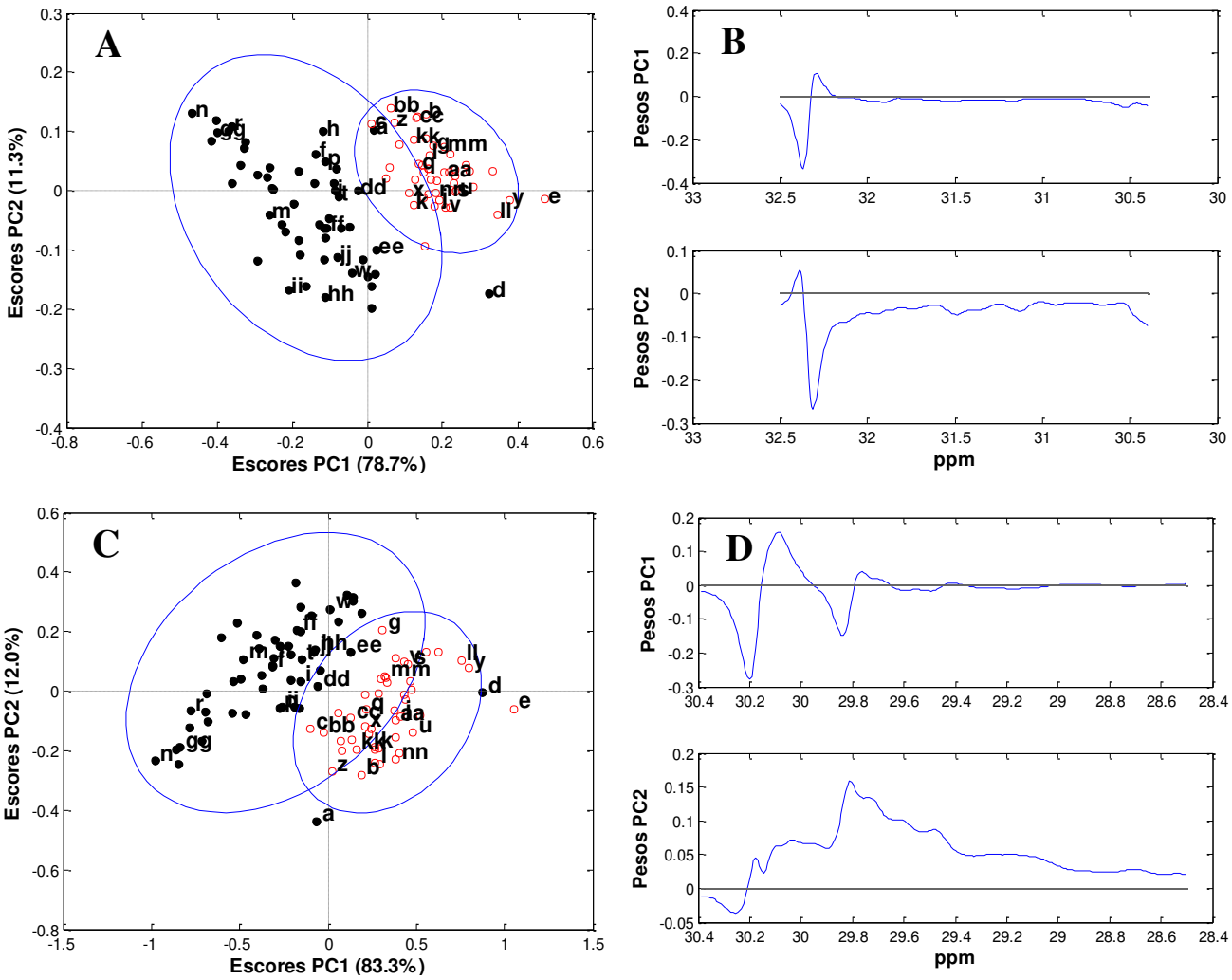

Figura 56: iPCA dos espectros de RMN ^{13}C. A e B, gráficos de escores e de pesos relativos ao intervalo 7; C e D, gráficos de escores e de pesos relativos ao intervalo 8. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

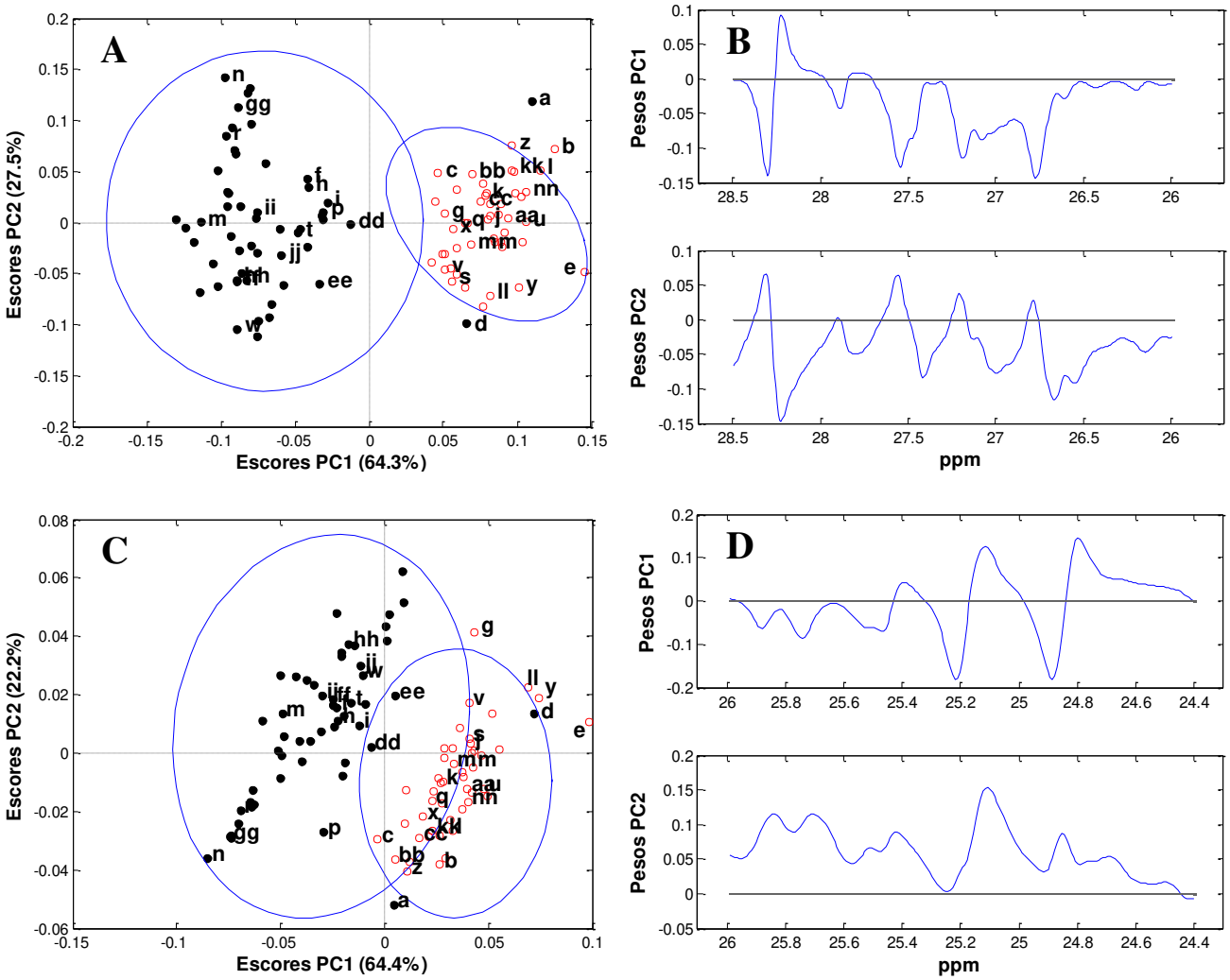

Figura 57: iPCA dos espectros de RMN ^{13}C. A e B, gráficos de escores e de pesos relativos ao intervalo 9; C e D, gráficos de escores e de pesos relativos ao intervalo 10. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

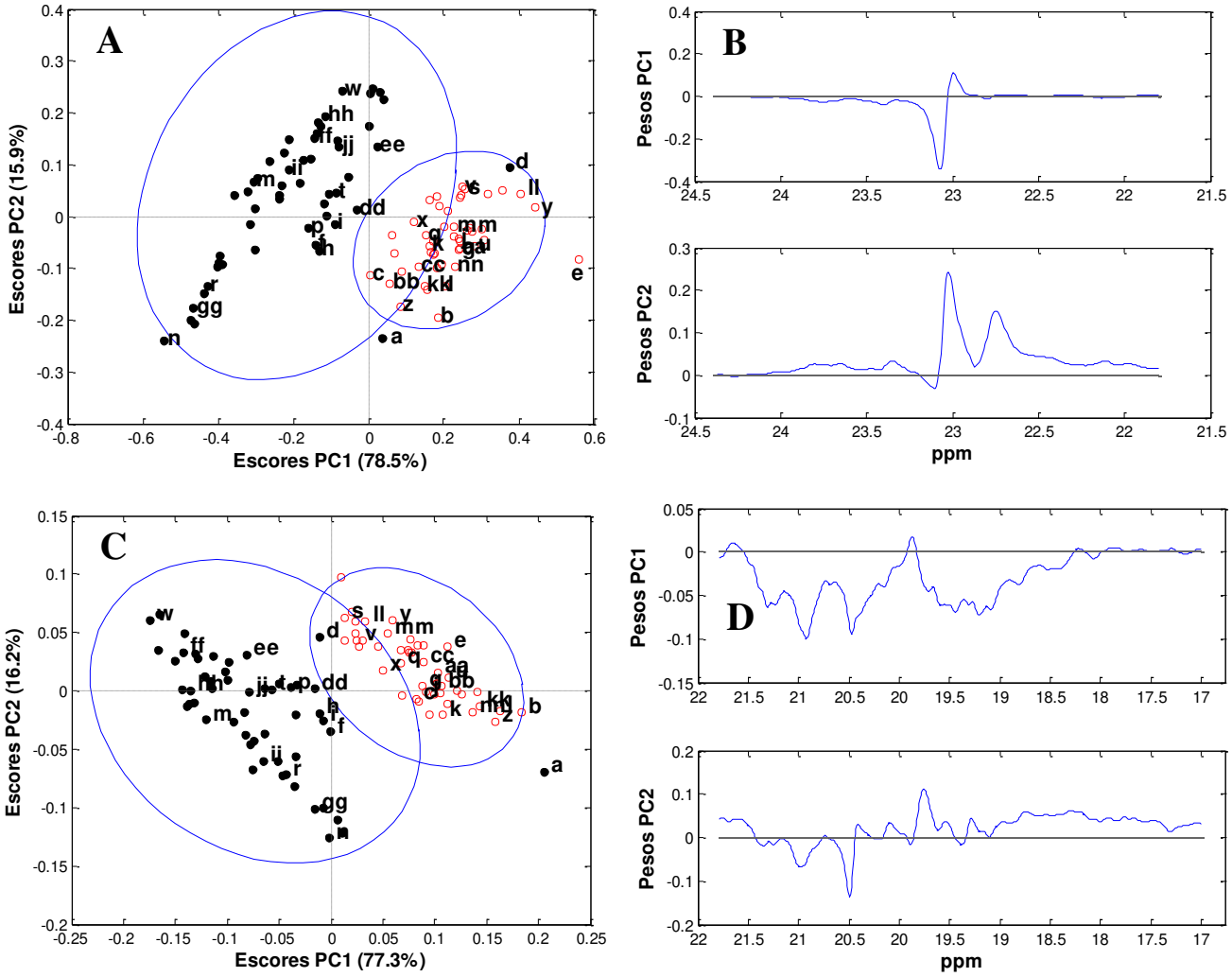

Figura 58: iPCA dos espectros de RMN ^{13}C. A e B, gráficos de escores e de pesos relativos ao intervalo 11; C e D, gráficos de escores e de pesos relativos ao intervalo 12. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

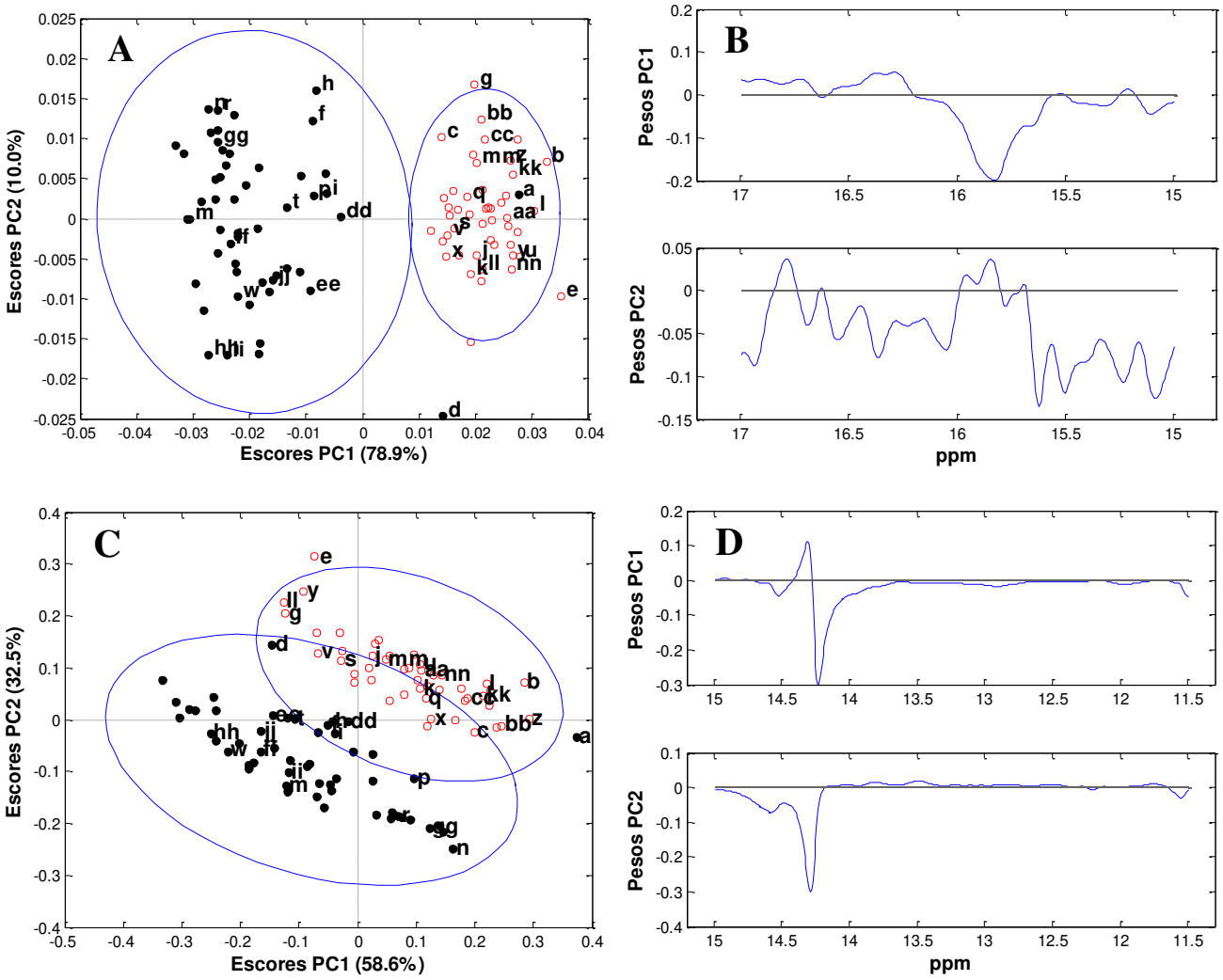

Figura 59: iPCA dos espectros de RMN ^{13}C. A e B, gráficos de escores e de pesos relativos ao intervalo 13; C e D, gráficos de escores e de pesos relativos ao intervalo 14. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

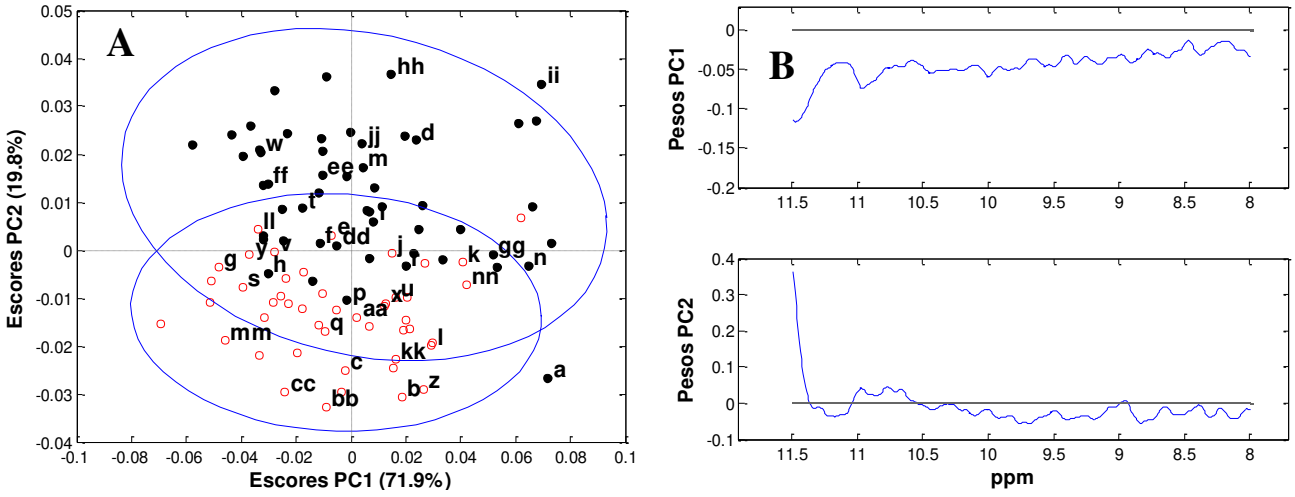

Figura 60: iPCA dos espectros de RMN ^{13}C. A e B, gráficos de escores e de pesos relativos ao intervalo 15. Amostras metropolitanas são identificadas por pontos pretos, enquanto as amostras não metropolitanas são identificadas por círculos vermelhos.

3.5.4. Considerações acerca dos experimentos de RMN

Conforme citado no item 3.4.2, os experimentos de RMN realizados neste trabalho não utilizaram a trava de frequência do campo magnético que necessita do uso de solventes deuterados, pois ela é feita exatamente usando o sinal de ressonância do deutério, como se fosse um "padrão interno", no intuito de corrigir pequenas variações no campo magnético.

Uma vez que os experimentos são feitos com o acúmulo de FIDs decorrentes de várias varreduras ou *scans* (um total de 16 para cada tipo de espectro, neste trabalho), a trava de frequência permite que a soma destes seja realizada apropriadamente, sempre entre os sinais de uma mesma frequência de ressonância. Caso haja significativas flutuações na frequência do campo magnético, há uma tendência para o alargamento dos sinais.

Neste trabalho, não foram detectados problemas devido à ausência do da trava de frequência, o que foi constatado, observando a largura de picos isolados (fora de

regiões "povoadas") nos dois tipos de espectros (RMN ^{1}H e ^{13}C). Para os espectros de RMN ^{1}H, acredita-se ainda que o procedimento de *bucketing* otimizado (capítulo 2) seja capaz de minimizar o alargamento moderado de picos, uma vez que a abordagem define uma região, em geral mais larga que os picos, para realizar a soma das intensidades. Também, vale ressaltar que o número de varreduras (*scans*) utilizado foi pequeno o que diminui a possibilidade da interferência das flutuações do campo magnético.

O uso de espectros sem trava de frequência pode ser considerado um ponto bastante positivo neste trabalho, pois com isto há uma diminuição no manuseio de amostra e uma economia quanto ao gasto de solventes deuterados.

Outra questão importante a ser considerada na aquisição dos espectros é o tempo de espera entre as varreduras, ou seja, o tempo entre os pulsos de excitação dos núcleos. Para uma análise de RMN quantitativa, um tempo de espera deve ser apropriado para que todos os núcleos relaxem (retornem ao estado de equilíbrio) antes de uma nova excitação e assim garanta-se que a cada FID somado, a intensidade máxima dos sinais seja adquirida. Existem vários fatores que influenciam no tempo e nos mecanismos de relaxação, porém, no contexto deste trabalho, pode ser citado apenas que quanto mais concentradas forem as amostras, menor é o tempo de relaxação dos núcleos [87].

As misturas de biodiesel-diesel foram investigadas puras, o que possibilitou baixos tempos de relaxação e baixos tempos na realização das análises. Além disso, como o tempo de relaxação depende do ambiente químico, do qual também depende a posição do sinal de ressonância e considerando o uso da análise de componentes principais em intervalos, pode-se considerar que se há uma perda de sinal, pela incompleta relaxação, esta deve ser mais ou menos a mesma para as regiões analisadas de deslocamentos químicos, não interferindo seriamente nas inferências. Por fim, é importante citar que a determinação do apropriado tempo de espera pode

ser realizada pelo experimento de recuperação de inversão (*inversion recovery*) [87], porém este não foi necessário neste trabalho.

No que diz respeito aos pulsos utilizados nos experimentos, também para uma análise de RMN quantitativa, é necessária a calibração do tempo do pulso para garantir que os núcleos sejam excitados apropriadamente (pulsos que promovam um giro de 90° na magnetização). Nesse contexto, o presente trabalho utilizou sequências de pulsos padrão, onde este parâmetro encontra-se otimizado.

3.6. Conclusões

A metodologia do presente trabalho mostrou-se muito útil para detectar amostras não conformes em misturas de biodiesel-diesel utilizando a análise de componentes principais em intervalos e a espectroscopia de ressonância magnética nuclear de hidrogênio, sendo ainda utilizada na correção dos desalinhamentos dos espectros a abordagem de *bucketing* otimizado. Observou-se que os intervalos estudados apontam diferentes amostras discrepantes e permitem a interpretação das características químicas das mesmas, o que por sua vez dá uma noção do perfil físico-químico da amostra, no que diz respeito aos parâmetros analisados no campo dos biocombustíveis. No entanto, um bom modelo para este fim, deve conter uma fração grande de amostras dentro das especificações, para garantir que as elipses de confiança, as quais servem como limite para a conformidade das amostras, sejam rigorosas o suficiente na detecção das amostras não conformes.

Neste trabalho, os modelos a partir dos espectros de RMN de hidrogênio permitiram a detecção de 39 amostras fora da especificação, sendo destas, 29 também reprovadas pelas análises físico-químicas para alguns parâmetros dos biocombustíveis. As outras 10 amostras "passaram" nestas análises, mas foram apontadas como discrepantes pelo método quimiométrico. O conjunto de dados

possuía 53 amostras reprovadas pela legislação Brasileira no que diz respeito aos parâmetros físico-químicos: teor de biodiesel, massa específica (densidade) a 20 °C, temperatura de destilação de 50 e 85% do volume da amostra, índice de cetano, teor de enxofre e ponto de fulgor. Deste modo, 24 amostras do conjunto de dados, reprovadas pelas análises citadas, principalmente devido ao teor de biodiesel (23 destas), não foram detectadas pela metodologia proposta envolvendo os espectros de RMN ^{1}H.

Acredita-se que amostras com teor de biodiesel ligeiramente fora da especificação sejam difíceis de serem detectadas devido à pequena diferença no padrão espectral para os sinais que envolvem essa discriminação. Se a especificação, que para as amostras estudadas é de 4,0 ± 0,5%, fosse, por exemplo, 4,0 ± 1,0%, permitindo que o padrão espectral entre amostras conformes e não conformes tivesse uma maior diferença, o número de amostras não detectadas com relação ao teor de biodiesel cairia para apenas 3. Assim, a porcentagem de acerto do método seria significativamente maior.

A detecção das 10 amostras que não haviam sido aprovadas pelos parâmetros físico-químicos disponíveis é considerada um ponto positivo desta metodologia, uma vez que é bastante provável que tais amostras sejam reprovadas por algum outro parâmetro que não foi analisado neste trabalho. Isto pode ser dito, baseando-se na distinção do padrão espectral destas amostras em alguns intervalos, o que apontou a possibilidade de problemas com as mesmas. Tais problemas podem ser adulterações, que em certas situações podem burlar até mesmo as análises de rotina no campo dos biocombustíveis. Como os espectros de RMN tendem a realizar uma investigação mais precisa da natureza química das amostras, acredita-se que a reprovação das amostras pelo modelo iPCA deve ser considerada relevante tanto quanto as reprovações pelas análises de rotina. Em muitos casos, o padrão espectral pode se mostrar mais poderoso do que estas últimas. Porém, vale ressaltar que dado o uso de

elipses com 95% de confiança, admite-se que uma em cada vinte amostras que deveriam ser encontradas dentro da elipse pode ser encontrada fora da mesma, representando, no contexto aqui estudado, um falso negativo. Assim, dentre as 10 amostras citadas acima algumas delas podem corresponder a este tipo de amostra, sendo isto inerente ao procedimento estatístico adotado nesta metodologia.

Claramente as amostras constituem dois tipos de populações, sendo as amostras metropolitanas com maior teor de olefinas, maior teor de *iso*-parafinas e mais ricas em compostos aromáticos mais simples (mononucleares) e as amostras não metropolitanas com maior teor de *n*-parafinas e de aromáticos polinucleares. Com as duas classes, é possível a exploração das suas diferenças na construção de um modelo com uma variabilidade maior ou explorar-se modelos para uma só classe, com variabilidade mais restrita. As duas opções são igualmente válidas e devem ser exploradas complementarmente.

A análise exploratória dos espectros de RMN ^{13}C mostrou-se mais complexa graças à etapa de redução de ruído. A análise de componentes principais em múltiplas escalas (MSPCA) mostrou uma boa filtragem de ruído, porém dependente da otimização de muitos passos. Apesar de ter sido feito de forma criteriosa, o MSPCA pode ter influenciado nos espectros finais e assim na iPCA para os mesmos. O desempenho destes modelos foi inferior àquele utilizando os espectros de hidrogênio, sendo detectadas 27 amostras do conjunto total, porém apenas 19 destas eram comprovadamente (através dos parâmetros físico-químicos) fora da especificação. As mesmas questões acerca do teor de biodiesel e das amostras não reprovadas detectadas pelo método quimiométrico podem ser ditas aqui.

Por fim, observou-se que a não utilização da trava de frequência ou *locking* do campo magnético, na aquisição dos espectros de RMN não promoveu sérios problemas às inferências e tendo em vista a diminuição no manuseio de amostra, consiste num ponto bastante positivo deste trabalho.

Capítulo 4: Escalamento de diferenças individuais multinível: uma nova abordagem para analisar a variação intra-individual de séries temporais em dados de ciências ômicas.

4.1. Introdução

As ciências ômicas são devotadas à: examinação de perfis metabólicos (metabolômica); investigação do conjunto de proteínas (proteômica); estudo da informação genética (genômica); de sistemas biológicos, enquanto eles são submetidos a, por exemplo, intervenção dietética, meio ambiente, estímulos patofisiológicos ou modificações genéticas. Um dos principais objetivos nas ciências ômicas é revelar biomarcadores discriminatórios que caracterizam mudanças nos padrões biológicos sobre diferentes grupos experimentais [106-109].

Atualmente, técnicas analíticas modernas de alta capacidade e alta-resolução, tais como, a cromatografia líquida acoplada à espectrometria de massas (CLEM), a cromatografia gasosa acoplada à espectrometria de massas (CGEM) e ressonância magnética nuclear, usadas para realizar as medidas nas ciências ômicas têm levado à aquisição de grandes quantidades de dados complexos. Tais medidas são frequentemente realizadas seguindo um planejamento experimental predefinido, envolvendo fatores como tempo, tratamento e dose [106-109].

Geralmente, conjuntos de dados adquiridos no curso do tempo nas ômicas podem ter conteúdos de informação multiníveis originados de múltiplas fontes de variação. As variações dos dados podem ter diferenças dinâmicas nos níveis das variáveis (metabólitos, genes, proteínas, etc.), associadas ao efeito do tratamento ou intervenção e variações devido às diferenças não dinâmicas ou estáticas, que podem estar presente mesmo antes da perturbação (tratamento). Estas últimas surgem das

abundâncias absolutas das variáveis que são características de cada indivíduo no estudo. Esta parte do conjunto de dados corresponde à variação que não é interessante no contexto do estudo dos efeitos do tratamento. De fato, a presença deste tipo de variação pode até mesmo dificultar as interpretações nas investigações ômicas, pois ela é muitas vezes maior do que as diferenças dinâmicas. As diferenças entre estes níveis absolutos estáticos, para cada indivíduo, refletem a chamada variação interindividual e as diferenças dinâmicas líquidas refletem a chamada variação intra-individual. Adicionalmente, sabe-se que os níveis de muitas variáveis, reveladas pelos conjuntos de dados, são altamente inter-relacionados através de rotas bioquímicas, que levam a uma maior complexidade [32,110-114].

A rede complicada de genes, proteínas ou metabólitos pode ser entendida visualizando os três paradigmas (Figura 61) relativos às diferenças entre dois grupos (por exemplo, um grupo de controle e um grupo de tratamento) num estudo onde há a medição de respostas em muitas variáveis [113]. A Figura 61A mostra as diferenças entre os valores de uma única destas variáveis para os dois grupos experimentais (investigação univariada), onde a correlação entre as variáveis é completamente desconsiderada e combinações em potencial não podem ser acessadas. Por outro lado, o paradigma mostrado na Figura 61B mostra a separação entre os grupos, que pode ser detectada devido à forte correlação entre duas variáveis (variáveis “X” e “Y”). Neste caso, métodos multivariados, tais como, análise de componentes principais (PCA) [115] ou análise discriminante por quadrados mínimos parciais (PLS-DA) [116], podem ser usados para acessar a separação entre os grupos. De fato, a procura por biomarcadores multivariados é muito popular nas ciências ômicas. Finalmente, a Figura 61C apresenta um paradigma onde nem métodos univariados, nem os métodos multivariados padrões podem evidenciar separações, pois as perturbações experimentais são seguidas de mudanças na estrutura de correlação. Os paradigmas

multivariados mostram que após um tratamento podem ocorrer efeitos comuns a todos os indivíduos (Figura 61B) ou efeitos distintos para os indivíduos (Figura 61C).

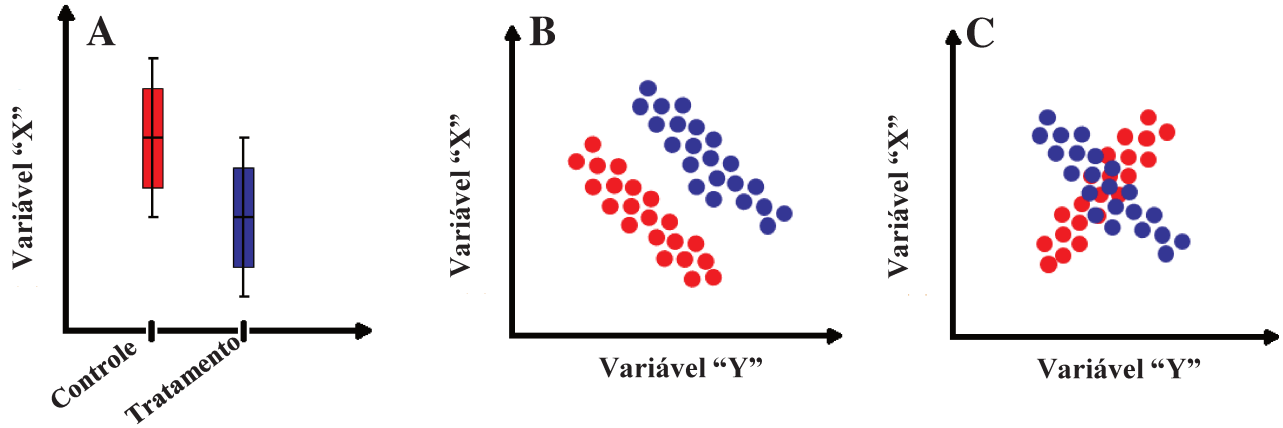

Figura 61: Três paradigmas para visualizar as diferenças entre dois grupos. **A:** diferenças nos níveis de um único biomarcador (variável). **B**: diferença nos níveis devido à correlação entre dois biomarcadores; efeitos comuns a todos os indivíduos. **C**: diferenças nos níveis com mudanças na estrutura de correlação; efeitos distintos para os indivíduos. Bolas vermelhas = grupo de controle e bolas azuis = grupo de tratamento.

Diante do exposto, uma abordagem holística para o tratamento de dados em ciências ômicas exige o foco: 1- em estrutura multiníveis, levando em conta a variação associada ao tratamento (variação intra-individual) e 2- na relação entre as variáveis (BVR, do inglês *Between Variables Relationship*) [113], objetivando considerar os paradigmas multivariados citados.

4.2. Objetivos

Neste capítulo, é apresentado um novo método chamado, escalamento de diferenças individuais multinível (ML-INDSCAL, do inglês *Multilevel Individual Differences Scaling*) que reúne as duas ideias citadas acima para a análise da variação intra-individual dos dados dinâmicos (séries temporais) em ciências ômicas. Portanto,

neste trabalho, objetiva-se explicar o método ML-INDSCAL, e as abordagens de onde o mesmo é derivado, tais como, o método de curvas de respostas principais (PRC, do inglês *Principal Response Curves*) [117], o método de trajetórias escaladas ao máximo, alinhadas e reduzidas (SMART, do inglês *Scaled-to-Maximum, Aligned, and Reduced Trajectories*) [118] e o método de escalamento de diferenças individuais (INDSCAL, do inglês *Individual Differences Scaling*) [113]. Dentro da apresentação do método é também explicada a estratégia *jack-knife* para validação dos modelos ML-INDSCAL, além da aplicação desta abordagem a um pequeno conjunto de dados simulado e a um conjunto de dados reais de um estudo de perfis de expressão de genes (genômica) celulares globais em linhas de células expressando a forma nativa da proteína Vpr e duas linhas com mutação (cepas R80A e F72A/R73A), disponível na literatura [119,120], comparando os resultados com os demais métodos.

4.3. Notações adicionais

Em adição às notações já definidas no capítulo 1, item 1.2, para este capítulo é necessário definir outras notações que serão utilizadas na apresentação da teoria sobre os métodos que seguem a ideia de separação de fontes de variação baseando-se no planejamento experimental. Os fatores de tratamento e tempo (ocasiões de medidas) são representados pelos índices (j = 1, …, J) e (k = 1, …, K), respectivamente. O índice (l = 1, …, L) se refere às variáveis (metabólitos, genes, proteínas, etc)., enquanto o índice (i = 1, ..., I) se refere ao número de indivíduos em cada grupo experimental, ou seja, existem I indivíduos submetidos a J tratamentos, dos quais medidas são tomadas em K ocasiões. Portanto, no estudo existem N (= IJK) amostras. A notação ponto nos subscritos indica sobre qual índice a média está sendo tomada (por exemplo, $x_{.jkl}$ é a média ao longo dos indivíduos no grupo j na ocasião k para a variável l). Ainda, sobre os comentários acerca dos planejamentos experimentais,

estes são considerados, em todos os casos, planejamentos balanceados, uma vez que cada grupo experimental possui o mesmo número de indivíduos.

4.4. Teoria

4.4.1. Planejamento experimental: exemplo

Um planejamento experimental diz respeito ao modo como uma determinada investigação é arranjada na busca por certa resposta experimental. O arranjo conta com a combinação de fatores que supostamente são relevantes para a determinação desta resposta, sendo também esta relevância muitas vezes testada. A resposta experimental pode ser um único valor (por exemplo, rendimento de um produto numa reação química, etc.) ou diversos valores (por exemplo, rendimentos de um produto principal e de um produto paralelo numa reação química, ou um espectro de RMN com L deslocamentos químicos arranjados num vetor, etc.) [121-124].

Neste trabalho, são de interesse os planejamentos experimentais com respostas que são usualmente vetores e por este motivo, um exemplo deste tipo é dado a seguir. Considere um planejamento experimental com replicação onde dois fatores, tempo e tratamento, são investigados. O número de tratamentos é quatro, sendo estes correspondentes a um grupo de controle, um grupo no qual foi administrado uma dose baixa de um medicamento, um grupo no qual foi administrado uma dose média de um medicamento e um grupo no qual foi administrado uma dose alta de um medicamento, assim $j = 1$ (controle), 2 (dose baixa), 3 (dose média), 4 (dose alta) e $J = 4$. Cada indivíduo é medido em três ocasiões, assim $k = 1$ (ocasião T1), 2 (ocasião T2), 3 (ocasião T3) e $K = 3$. O número de replicatas (indivíduos) é três por grupo de tratamento por ocasião de medida, assim $i_{jk} = 1, ..., I$, sendo $I = 3$. Suponha também que a medida corresponde à aquisição de espectros de RMN ^{1}H a partir de algum

fluido biológico (urina, plasma, etc.) dos indivíduos. Os espectros possuem todos L deslocamentos químicos e, portanto, são vetores linha $\mathbf{x}^T$ ($1 \times L$). Desse modo existem N (= IJK) amostras, ou seja, 36 (= 3*4*3) amostras no planejamento experimental. A Figura 62 mostra o planejamento do experimento para um grupo de tratamento [125].

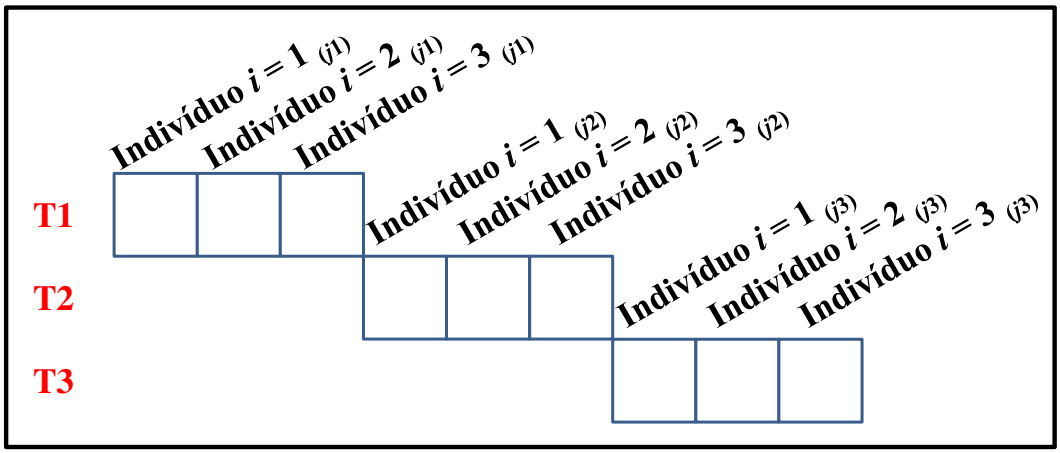

Figura 62: Planejamento do experimento dado como exemplo. Cada quadrado corresponde a um espectro de RMN ^{1}H, ou seja, uma amostra. A figura representa o planejamento de um grupo de tratamento, assim note que o experimento total consiste de quatro destes grupos de quadrados (controle, dose baixa, dose média e dose alta).

Cada quadrado na Figura 62 representa uma amostra e, portanto, um espectro de RMN. Observe que há a representação de apenas um grupo de tratamento que pode ser arranjado numa submatriz $\mathbf{X}_j$ ($IK \times L$) = $\mathbf{X}_j$ ($9 \times L$), e o conjunto de dados inteiro é composto por quatro destas submatrizes (quatro grupos de tratamento ou quatro dos grupos de quadrados na Figura 62). Deste modo, o conjunto inteiro é $\mathbf{X}$ ($IJK \times L$) = $\mathbf{X}$ ($N \times L$) = $\mathbf{X}$ ($36 \times L$). A partir desta última matriz, métodos específicos podem ser utilizados para estudar as fontes de variação devido a cada fator do planejamento experimental (tratamento e tempo, neste caso), além da fonte de variação devido à interação dos fatores. Tais métodos serão comentados a seguir.

4.4.2. Métodos baseados na separação das fontes de variância

Nas ciências ômicas (genômica, metabolômica, transcriptômica, proteômica, etc.), rotineiramente, centenas ou milhares de variáveis são medidas simultaneamente sobre indivíduos, que por sua vez estão submetidos a algum planejamento experimental. Os efeitos dos fatores experimentais tais como, tempo e tratamento, podem ser estimados para cada variável dependente utilizando a análise de variância (ANOVA, do inglês *Analysis of Variance*) [121] como é tipicamente feito quando uma simples variável dependente (resposta experimental) é obtida. No entanto, tal análise não considera as relações entre as variáveis e desse modo não é adequada para analisar os dados em ômicas. A extensão multivariada da ANOVA, denominada ANOVA multivariada (MANOVA, do inglês *Multivariate* ANOVA) [122] também não é adequada para os dados citados, pois este método exige que o número de amostras (indivíduos) seja menor do que o número de variáveis (em outras palavras, a matriz de variância-covariância precisa ter posto completo para que seja invertível, pois esta matriz nesta forma é usada para testes de significância), além de pressupor uma distribuição multinormal dos dados, o que, geralmente, não é o caso para os dados nas ciências ômicas [123].

Valendo-se do planejamento experimental, outros métodos são sugeridos na literatura [124] para contornar as limitações da MANOVA. Tais métodos estendem as ideias da ANOVA, separando as fontes de variação e aplicando análises de componentes em matrizes específicas. Aqui cinco destes métodos serão comentados, a saber, a análise de variância – análise de componentes simultâneas (ANOVA-SCA ou ASCA, do inglês ANOVA *Simultaneous Component Analysis*) [125-129], a análise de componentes simultâneas multinível (MSCA, do inglês *Multilevel Simultaneous Component Analysis*) [112], a análise de variância – análise de componentes

principais (ANOVA-PCA ou APCA, do inglês ANOVA *Principal Component Analysis*) [130], e os métodos PRC e SMART.

ASCA permite a análise e a interpretação separada de modelos para cada uma das diferentes fontes de variações dos dados, induzidas por diferentes fatores do planejamento experimental, sendo que até mesmo as interações entre os fatores podem ser estudadas. O método se baseia no modelo linear mostrado na Equação 28:

$$x_{ijkl} = \mu_l + \alpha_{jl} + \beta_{kl} + (\alpha\beta)_{jkl} + e_{ijkl} \qquad \text{Equação 28}$$

Onde x_{ijkl} é o valor da variável l no tratamento j e tempo k para o indivíduo i; α_{jl} representa o efeito do fator tratamento; β_{kl} representa o efeito do fator tempo; $(\alpha\beta)_{jkl}$ representa a interação entre o tratamento e o tempo; e_{ijkl} representa o erro associado à estimativa dos efeitos. O modelo linear assumido acima possui infinitas respostas e por este motivo são impostas as restrições usuais dos modelos ANOVA (Equação 29) para permitir uma modelagem, com uma única solução, da variação de uma variável individual, como mostrado na Equação 30.

$$\sum_{j}^{J} \alpha_{jl} = 0(\forall l); \sum_{k}^{K} \beta_{kl} = 0(\forall l);$$
$$\sum_{j}^{J} (\alpha\beta)_{jkl} = 0(\forall k, l); \sum_{k}^{K} (\alpha\beta)_{jkl} = 0(\forall j, l) \qquad \text{Equação 29}$$

$$x_{ijkl} = x_{\dots l} + \left(x_{.j.l} - x_{\dots l}\right) + \left(x_{..kl} - x_{\dots l}\right) + \left(x_{.jkl} - x_{.j.l} - x_{..kl} + x_{\dots l}\right) + \left(x_{ijkl} - x_{.jkl}\right) \qquad \text{Equação 30}$$

Os termos do lado direito da Equação 30 são, respectivamente, a média global sobre todas as amostras, o efeito principal do fator tratamento, o efeito principal do fator tempo, a interação tratamento *versus* tempo, e a contribuição individual. Em

outras palavras, as variáveis são particionadas em uma soma de médias (variação sistemática) e uma contribuição individual. Cada uma das contribuições para a variância é coletada em uma matriz apropriada, como mostrado na Equação 31:

$$\mathbf{X} = \mathbf{X}_M + \mathbf{X}_J + \mathbf{X}_K + \mathbf{X}_{JK} + \mathbf{X}_{IND} \qquad \text{Equação 31}$$

Onde todas as matrizes têm dimensão ($N \times L$) e possuem nas suas linhas a contribuição por amostra de cada uma das fontes de variação, na mesma ordem mostrada na Equação 30. Pode ser demonstrado que os espaços coluna das matrizes são ortogonais entre si [128], o que permite a partição (em partes independentes) da variância em somas de quadrados, como mostrado na Equação 32, que pode ser usada para determinar a contribuição de cada fator e interação na variância total dos dados.

$$\|\mathbf{X}\|^2 = \|\mathbf{X}_M\|^2 + \left\|\mathbf{X}_J\right\|^2 + \|\mathbf{X}_K\|^2 + \left\|\mathbf{X}_{JK}\right\|^2 + \|\mathbf{X}_{IND}\|^2 \qquad \text{Equação 32}$$

O próximo passo na ASCA é o uso de modelos de componentes SCA para aproximar as informações contidas nas quatro últimas matrizes do lado direito da Equação 31. Existem diferentes versões de SCA, mas em ASCA a chamada SCA-P é utilizada, onde modelos PCAs comuns (capítulo 1) são obtidos para cada matriz citada. Deste modo, as variações associadas a cada fator são analisadas sem serem confundidas, em um subespaço de dimensão reduzida próprio, cujas amostras são projetadas. As interpretações dos submodelos são feitas de modo mutualmente independente, dada a ortogonalidade dos espaços coluna das matrizes iniciais, o que garante que os espaços coluna dos submodelos sejam também ortogonais [128]. A Equação 33 mostra os modelos SCA-P em ASCA, onde as entidades **T** ($N \times R$) e **P** ($L \times R$) são as matrizes de escores e pesos obtidas para R componentes principais, sendo R dependente do submodelo, ou seja, ele pode ser R_J, R_K, R_{JK} ou R_{IND}; o vetor **m** ($L \times$

1) contém as médias das colunas de $\mathbf{X}$; a entidade $\mathbf{1}$ ($N \times 1$) é um vetor de uns; e a matriz $\mathbf{E}$ ($N \times L$) contém os resíduos dos modelos.

$$\mathbf{X} = \mathbf{1}\mathbf{m}^{\mathrm{T}} + \mathbf{T}_J\mathbf{P}_J^{\mathrm{T}} + \mathbf{T}_K\mathbf{P}_K^{\mathrm{T}} + \mathbf{T}_{JK}\mathbf{P}_{JK}^{\mathrm{T}} + \mathbf{T}_{IND}\mathbf{P}_{IND}^{\mathrm{T}} + \mathbf{E} \qquad \text{Equação 33}$$

Apesar de a ASCA ter sido comentada aqui para um planejamento experimental particular, as ideias podem ser estendidas para diferentes planejamentos e modelos lineares ANOVA, inclusive com um número maior de fatores envolvidos [124].

Um caso especial da ASCA é a MSCA, particularmente útil para a investigação de dados temporais em estudos cruzados (*crossover design*). Quando um único fator experimental é avaliado, a MSCA decompõe os dados numa média global, uma parte relativa à variação interindividual e outra relativa à variação intra-individual. Para a exposição matemática da MSCA, considere que a partir de I indivíduos (i = 1, ..., I) são medidas L variáveis (l = 1, …, L) em K ocasiões (k = 1, …, K), assim cada indivíduo tem K amostras associadas a ele e que compõem matrizes $\mathbf{X}_i$ ($K \times L$) que quando concatenadas verticalmente resultam na matriz $\mathbf{X}$ ($N \times L$), onde $N = IK$. Inicialmente a matriz $\mathbf{X}$ é centrada na média fornecendo a matriz $\mathbf{X}_{\mathrm{CM}}$. A partir desta última matriz, cada matriz $\mathbf{X}_{\mathrm{CM}i}$ ($K \times L$) é isolada e centrada na média individualmente gerando as matrizes $\mathbf{X}_{\mathrm{CMCM}i}$ ($K \times L$), cuja concatenação vertical fornece a matriz $\mathbf{X}_{\mathrm{C}}$ ($N \times L$) contendo a variação intra-individual. As médias $\mathbf{m}_i$ ($L \times 1$) de cada matriz $\mathbf{X}_{\mathrm{CM}i}$ são concatenadas para compor a matriz $\mathbf{M}$ ($I \times L$) contendo a variação interindividual. As etapas de centragem são mostradas na Equação 34:

$$\begin{aligned} \mathbf{X}_{\mathrm{CM}} &= \mathbf{X} - \mathbf{1}_N\mathbf{m}^{\mathrm{T}} \\ \mathbf{X}_{\mathrm{CMCM}i} &= \mathbf{X}_{\mathrm{CM}i} - \mathbf{1}_K\mathbf{m}_i^{\mathrm{T}} \end{aligned} \qquad \text{Equação 34}$$

Onde $\mathbf{m}$ ($L \times 1$) é um vetor com as médias das colunas de $\mathbf{X}$; $\mathbf{1}_N$ é um vetor de uns com dimensão ($N \times 1$); e $\mathbf{1}_K$ é um vetor de uns com dimensão ($K \times 1$).

Os submodelos para as variações interindividual e intra-individual são, portanto, obtidos a partir da PCA das matrizes $\mathbf{M}$ e $\mathbf{X}_C$, respectivamente, como mostrado na Equação 35:

$$\begin{aligned} \mathbf{M} &= \mathbf{T}_B\mathbf{P}_B^T + \mathbf{E}_B \\ \mathbf{X}_C &= \mathbf{T}_W\mathbf{P}_W^T + \mathbf{E}_W \end{aligned} \qquad \text{Equação 35}$$

Onde $\mathbf{T}_B$ ($I \times R_B$) e $\mathbf{P}_B$ ($L \times R_B$) são as matrizes de escores e pesos para a variação interindividual que é aproximada num subespaço com R_B dimensões (componentes principais) gerando os resíduos na matriz $\mathbf{E}_B$ ($I \times L$); e $\mathbf{T}_W$ ($N \times R_W$) e $\mathbf{P}_W$ ($L \times R_W$) são as matrizes de escores e pesos para a variação intra-individual que é aproximada num subespaço com R_W componentes principais gerando os resíduos na matriz $\mathbf{E}_W$ ($N \times L$). MSCA, a exemplo da ASCA, permite a interpretação de modelos independentes sem confundir as variações intra- e interindividual. Mais detalhes sobre a MSCA podem ser encontrados na literatura [112].

A ANOVA-PCA [130] é similar a ASCA, sendo ambos baseados na separação das fontes de variação seguindo um modelo linear ANOVA. A diferença entre os dois métodos é que ASCA usa a PCA sobre cada uma das matrizes dos efeitos (Equação 33), enquanto a ANOVA-PCA primeiramente adiciona a matriz com os erros da ANOVA (aqueles coletados na matriz $\mathbf{X}_{IND}$ na Equação 31) às matrizes dos efeitos e então aplica a PCA na soma resultante. Consequentemente, em ASCA apenas as relações entre os níveis médios das variáveis podem ser acessadas, o que de certa forma limita a quantidade de informação obtida e a significância do modelo. Na ANOVA-PCA o agrupamento das replicatas (indivíduos) pode ser obtido diretamente do modelo a cada nível dos efeitos. Apesar desta limitação da ASCA, na literatura já existe uma proposição para contorná-la, consistindo na projeção dos resíduos contidos na matriz $\mathbf{X}_{IND}$ nos subespaços gerados nos modelos dos efeitos para acessar as informações sobre os agrupamentos das amostras [123].

PRC [117] e SMART [118] são outros dois métodos que seguem a ideia de separação de fontes de variação baseando-se no planejamento experimental, estendendo ideias da análise univariada (análise de variância - ANOVA) e então aplicando PCA em matrizes específicas que carregam a variação associada à questão biológica em estudo [124].

O método PRC foi proposto para a análise de dados a partir de experimentos planejados com repetição no tempo, quando há um grupo de controle. O método expressa toda a variação nos dados relativa ao grupo de controle, permitindo o foco no efeito do tratamento que é dependente do tempo. Portanto, a variação intra-individual é evidenciada tomando os níveis no grupo de controle como a variação estática (não dinâmica). A Equação 36 mostra o modelo linear (o grupo controle é tomado como $j = 1$) adotado como base para o PRC para um planejamento experimental simples: dados balanceados (mesmo número de indivíduos em cada grupo experimental) e dois fatores, tratamento ($j = 1, \ldots, J$) e tempo ($k = 1, \ldots, K$).

$$x_{ijkl} = \alpha_{1kl} + (\alpha\beta)_{jkl} + e_{ijkl} \qquad \text{Equação 36}$$

Onde x_{ijkl} é o nível da variável l no tratamento j e tempo k para o indivíduo i; α_{1kl} é o nível médio da variável l no tempo k no grupo controle $j = 1$; $(\alpha\beta)_{jkl}$ representa o efeito específico como uma interação dos fatores tratamento e tempo; e e_{ijkl} representa o erro associado à estimativa.

Em contraste ao PRC, o método SMART é baseado em outra ideia para a separação das fontes de variação baseando-se no planejamento experimental. Neste, após uma etapa especial de escalamento (que não será comentada neste trabalho), os efeitos são tomados relativos ao primeiro ponto no tempo, objetivando remover as diferenças na pré-dose como diferenças grosseiras de magnitude nos dados. Neste

caso, o modelo linear (o primeiro ponto é tomado como k = 1) usando o mesmo planejamento experimental citado acima é:

$$x_{ijkl} = \beta_{j1l} + (\alpha\beta)_{jkl} + e_{ijkl} \qquad \text{Equação 37}$$

Onde podem ser vistos os mesmo símbolos usados para a Equação 36, exceto para β_{j1l} que é o nível médio da variável *l* no grupo *j* e ponto do tempo k = 1.

Para a estimativa dos efeitos, os modelos PRC e SMART são sujeitos às restrições mostradas na Equação 29 (restrições usuais da ANOVA) e a restrições adicionais, mostradas na Equação 38:

$$\begin{gathered}(\alpha\beta)_{1kl} = 0(\forall k, l);\ \text{PRC} \\ (\alpha\beta)_{j1l} = 0(\forall j, l);\ \text{SMART}\end{gathered} \qquad \text{Equação 38}$$

As decomposições resultantes para PRC e SMART sob as restrições são mostradas nas Equações 39 e 40, respectivamente. Uma discussão sobre as restrições, a estimativa dos efeitos e outras propriedades dos métodos PRC e SMART pode ser encontrada na literatura [117,118,124].

$$x_{ijkl} = x_{.1kl} + \left(x_{.jkl} - x_{.1kl}\right) + \left(x_{ijkl} - x_{.jkl}\right);\ \text{PRC} \qquad \text{Equação 39}$$

$$x_{ijkl} = x_{.j1l} + \left(x_{.jkl} - x_{.j1l}\right) + \left(x_{ijkl} - x_{.jkl}\right);\ \text{SMART} \qquad \text{Equação 40}$$

Após a decomposição, os termos na direita das Equações 39 e 40 são coletados em matrizes apropriadas e em especial, o segundo termo do lado direito $\left(x_{.jkl} - x_{.1kl}\right)$ e $\left(x_{.jkl} - x_{.j1l}\right)$ que são o alvo do estudo, são coletados nas matrizes $\mathbf{X}_{\text{PRC}}$ ($IJK \times L$) e $\mathbf{X}_{\text{SMART}}$ ($IJK \times L$), respectivamente. Em ambos, PRC e SMART, estas matrizes são aproximadas em um modelo PCA, fornecendo a descrição do fenômeno em um espaço de dimensão reduzida, como mostrado nas Equações 41 e 42:

$$\mathbf{X}_{\mathrm{PRC}} = \mathbf{T}_{\mathrm{PRC}}\mathbf{P}_{\mathrm{PRC}}^{\mathrm{T}} + \mathbf{E}_{\mathrm{PRC}} \quad \text{Equação 41}$$

$$\mathbf{X}_{\mathrm{SMART}} = \mathbf{T}_{\mathrm{SMART}}\mathbf{P}_{\mathrm{SMART}}^{\mathrm{T}} + \mathbf{E}_{\mathrm{SMART}} \quad \text{Equação 42}$$

Onde $\mathbf{T}_{\mathrm{PRC}}$ ($N \times R_{\mathrm{PRC}}$) e $\mathbf{P}_{\mathrm{PRC}}$ ($L \times R_{\mathrm{PRC}}$) são as matrizes de escores e pesos para a variação interindividual, segundo a partição em PRC, que é aproximada num subespaço com R_{PRC} dimensões (componentes principais) gerando os resíduos na matriz $\mathbf{E}_{\mathrm{PRC}}$ ($N \times L$), sendo N o número total de amostras; e $\mathbf{T}_{\mathrm{SMART}}$ ($N \times R_{\mathrm{SMART}}$) e $\mathbf{P}_{\mathrm{SMART}}$ ($L \times R_{\mathrm{SMART}}$) são as matrizes de escores e pesos para a variação interindividual, segundo a partição em SMART, que é aproximada num subespaço com R_{SMART} dimensões (componentes principais) gerando os resíduos na matriz $\mathbf{E}_{\mathrm{SMART}}$ ($N \times L$).

Diante do exposto, observa-se que diferentes métodos surgem devido a distintos modelos lineares ANOVA, ou seja, dependendo da maneira como a partição na variância é feita. A escolha de um método irá, portanto, depender da questão biológica em estudo, além do planejamento experimental executado. Considerando, por exemplo, os métodos PRC e SMART, dificilmente as matrizes $\mathbf{T}_{\mathrm{PRC}}$ e $\mathbf{T}_{\mathrm{SMART}}$ serão iguais, mesmo usando um mesmo número de componentes principais, ou seja, não é garantido que as matrizes de pesos gerem os mesmos subespaços e nem que transformações (rotações, por exemplo) em uma levem à outra. Assim, pode-se concluir que os métodos são diferentes formas de visualizar o conjunto de dados, devendo ser utilizados da forma mais objetiva possível.

4.4.3. Escalamento de Diferenças Individuais (INDSCAL)

INDSCAL é uma abordagem proposta por Carroll e Chang [131,132] útil para investigação de relações entre matrizes simétricas de similaridade. O método

INDSCAL corresponde a uma versão restrita da abordagem CANDECOMP/PARAFAC [133] para arranjos com "fatias" simétricas (como as matrizes de covariância), onde as fatias compartilham uma matriz de pesos comum em um subespaço de dimensão reduzida. Cada dimensão do subespaço é diferentemente ponderada por diferentes indivíduos (neste contexto, os indivíduos são os grupos experimentais representados por suas respectivas matrizes de covariância), sendo a magnitude da ponderação contida numa matriz diagonal com elementos não negativos.

Na literatura recente [113], a abordagem INDSCAL foi sugerida como um método capaz de evidenciar as relações entre os metabólitos (BMR, do inglês *Between Metabolites Relationship*) para dados em metabolômica. Neste trabalho, modifica-se este acrônimo para BVR (do inglês *Between Variable Relationship*) para torná-lo mais geral para as ciências ômicas. As relações citadas podem ser evidenciadas operando-se diretamente num arranjo de três modos $\underline{\mathbf{R}}$ composto pelas matrizes de covariância, que são as fatias do arranjo calculadas para cada ocasião experimental. Por exemplo, se a matriz $\mathbf{X}_k$ ($I \times L$) com I indivíduos e L variáveis é relacionada a um ponto no tempo (ocasião experimental k), então a matriz de covariância $\mathbf{R}_k$ de tamanho ($L \times L$) pode ser computada de acordo com a Equação 43:

$$\mathbf{R}_k = \frac{1}{I-1}\left(\mathbf{X}_k - \mathbf{1}_I \mathbf{m}_k^{\mathrm{T}}\right)^{\mathrm{T}}\left(\mathbf{X}_k - \mathbf{1}_I \mathbf{m}_k^{\mathrm{T}}\right) \qquad \text{Equação 43}$$

Onde $\mathbf{1}_I$ é um vetor de uns, com comprimento I e $\mathbf{m}_k$ ($1 \times L$) é um vetor com as médias das colunas de $\mathbf{X}_k$.

O modelo assumido em INDSCAL é mostrado na Equação 44, em que a função f é minimizada num sentido de quadrados mínimos, encontrando valores ótimos para $\mathbf{A}$ e $\mathbf{G}_K$:

$$\text{Modelo: } \mathbf{R}_k = \mathbf{A}\mathbf{G}_k\mathbf{A}^{\mathrm{T}} + \mathbf{E}_k$$

$$\text{Minimização: } f(\mathbf{A}, \mathbf{G}_k) = \sum_{k=1}^{K} \left\| \mathbf{R}_k - \mathbf{A}\mathbf{G}_k\mathbf{A}^{\mathrm{T}} \right\|^2$$

Restrições: $\mathbf{G}_k$ = matriz diagonal com elementos não negativos;

$$\text{Ortogonalidade: } (\mathbf{A}^{\mathrm{T}}\mathbf{A}) = \mathbf{I}_R;$$

$$\text{Sem ortogonalidade: } diagonal\ de\ (\mathbf{A}^{\mathrm{T}}\mathbf{A}) = \mathbf{1}_R$$

Equação 44

Onde R é o número de componentes usados no modelo; $\mathbf{G}_k$ ($R \times R$) contém as ponderações das ocasiões k na sua diagonal; $\mathbf{A}$ é a matriz de pesos ($L \times R$); $\mathbf{R}_k$ é a matriz de covariância ($L \times L$) associada à ocasião k (fatias no arranjo $\underline{\mathbf{R}}$); $\mathbf{E}_k$ ($L \times L$) contém os resíduos; e $\mathbf{I}_R$ é uma matriz identidade ($R \times R$).

Dois tipos de restrições sobre a matriz de pesos podem ser impostos, mas em termos práticos a solução usando uma ou outra restrição é virtualmente a mesma, quando o requerimento de simetria para a matriz de pesos é satisfeita. Entretanto, o uso da restrição de ortogonalidade na matriz de pesos tem sido sugerido para contornar algumas dificuldades, tais como, soluções não simétricas, ponderações negativas, e o problema da degenerescência (obtenção de soluções que não são mínimos globais e que apresentam altos valores de ponderações associados à matriz **A** com colunas altamente correlacionadas) [132]. Portanto, neste trabalho, escolheu-se usar a versão do INDSCAL com restrição de ortogonalidade. Discussões detalhada sobre estas restrições e outras propriedades do método INDSCAL podem ser encontradas na literatura [131-134].

4.4.4. Escalamento de diferenças individuais multinível (ML-INDSCAL)

Apesar das vantagens das abordagens ASCA, ANOVA-PCA, MSCA, PRC e SMART para descrição separada da variância dos dados, estes métodos não são

capazes de evidenciar completamente as relações entre as variáveis (BVRs). Em outras palavras, tais métodos, por suas características, não focam nas mudanças de correlação ou de covariância entre as variáveis em cada grupo. Assim, essas mudanças permanecem escondidas e pontos detalhados sobre as BVRs não são considerados. Para resolver este problema é necessário calcular as matrizes de covariância de cada grupo experimental e investigar as mudanças comuns nestas relações. Isto pode ser feito usando o método INDSCAL. Vale ressaltar que as covariâncias e variâncias são preferidas frente às correlações (coeficientes de Pearson), pois, estas últimas podem omitir os aspectos quantitativos, uma vez que escalam os níveis das variáveis pelos desvios padrões [113].

Neste contexto, propõe-se o método chamado, escalamento de diferenças individuais multinível (ML-INDSCAL) para investigar os dados resolvidos no tempo, reunindo o foco na variação intra-individual e nas mudanças de estrutura de covariância, considerando em detalhes as BVRs. Como no INDSCAL convencional, a abordagem ML-INDSCAL é aplicada num arranjo composto pelas matrizes de covariância de cada grupo experimental. Entretanto, antes da montagem do arranjo, é realizada uma separação da variância com o objetivo de isolar a variação intra-individual da variação interindividual no conjunto de dados. Esta separação pode ser feita seguindo procedimentos similares àqueles dos métodos PRC ou SMART, ou seja, expressando a variação dos dados em relação ao grupo de controle (grupo placebo) ou ao primeiro ponto de tempo, respectivamente. A escolha por um ou outro procedimento irá depender do planejamento experimental usado para a aquisição dos dados ou à questão biológica do estudo.

As Figuras 63 e 64 apresentam os procedimentos esquemáticos para isolar a variação intra-individual no método ML-INDSCAL. A matriz subtraída (grupo controle ou primeira ocasião de medida) pode ser composta diferentemente, dependendo das características do planejamento experimental. Se o experimento é

feito de maneira que cada indivíduo é submetido a todos os tratamentos (planejamento cruzado ou *crossover design*), as medidas obtidas no grupo controle são subtraídas diretamente indivíduo por indivíduo, sendo assim consideradas como um "branco" ou a variação não dinâmica. Por outro lado, se o experimento é realizado com diferentes grupos de indivíduos em cada tratamento (planejamento paralelo ou *parallel design*), a média do grupo controle pode ser subtraída de cada um dos demais grupos, sendo esta a variação não dinâmica considerada. Em ambos os casos a ideia de separação da variância similar ao PRC é usada (Figura 63). Alternativamente, a ideia de partição seguida por SMART pode ser aplicada (Figura 64), sendo nos dois planejamentos citados, as medidas relativas ao primeiro ponto do tempo usadas como variação não dinâmica.

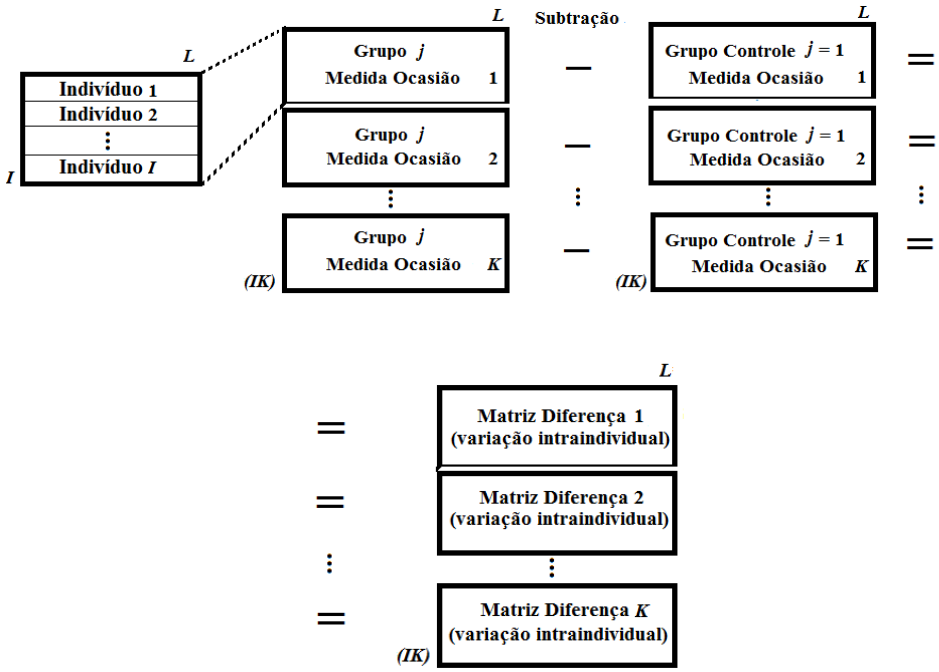

Figura 63: Estrutura do conjunto de dados e separação da variância similar ao método PRC. *K* é o número total de ocasiões (pontos de tempo) no estudo. *I* é o número de indivíduos em cada ocasião e *L* é o número de metabólitos (variáveis) no conjunto de dados. Os grupos de tratamento são representados por $j = 2, \ldots, J$ e o grupo controle por $j = 1$. Neste exemplo é suposto um planejamento experimental balanceado.

Espera-se que o ML-INDSCAL forneça modelos sem viés (*unbiased models*) pelo uso dos sistemas biológicos atuando como seus próprios controles, o que pode contornar os problemas com amostras (indivíduos) discrepantes (*outliers*), que por sua vez ocorrem devido às variações biológicas estáticas de alguns indivíduos que são muito diferentes da média do seu grupo. A subtração desta variação é feita usando a relação dos dados com o grupo controle ou com a primeira ocasião de medida. Isto é uma vantagem em relação ao método INDSCAL convencional [113].

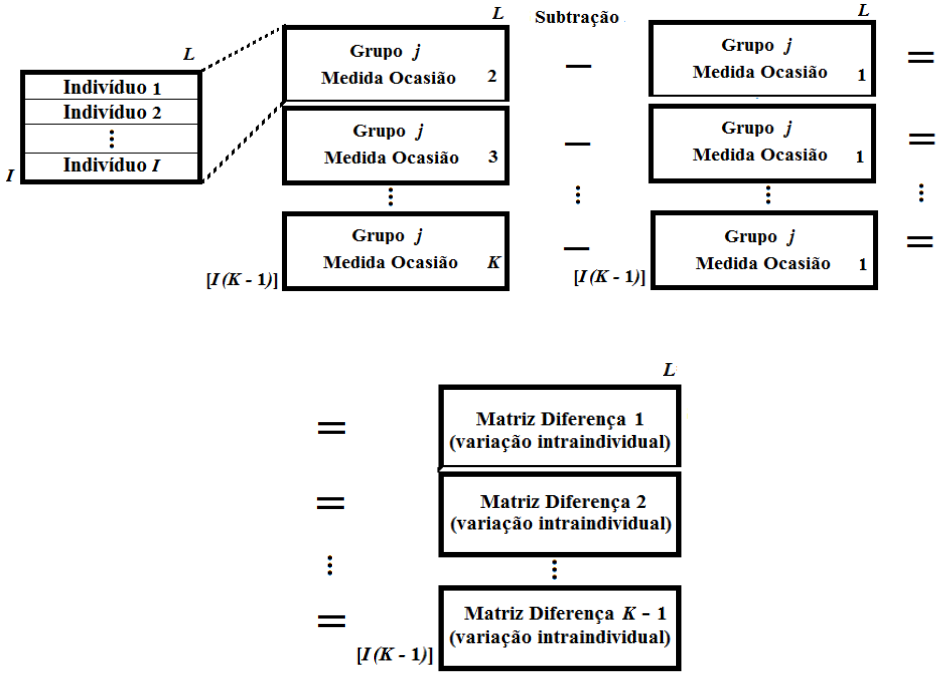

Figura 64: Estrutura do conjunto de dados e separação da variância similar ao método SMART. *K* é o número total de ocasiões (pontos de tempo) no estudo. *I* é o número de indivíduos em cada ocasião e *L* é o número de metabólitos (variáveis) no conjunto de dados. Os grupos de controle e tratamento são representados por $j = 1, \ldots, J$. Neste exemplo é suposto um planejamento experimental balanceado.

Como observado nas Figuras 63 e 64, as separações resultam em matrizes com diferenças, de onde as matrizes de covariância são calculadas e assim organizadas em um arranjo de três modos (Figura 65). Como mostrado na Figura 65, cada grupo de tratamento *j* irá fornecer um arranjo de três modos qualquer que seja a partição

seguida (similar a PRC ou SMART). Se o tratamento $j = 1, \ldots, J$ representa diferentes doses (e.g., doses de uma droga) no planejamento experimental, cada arranjo de três modos **R** (K ou K-1 × L × L) de cada grupo experimental (tratamento) será organizado para compor um arranjo maior **R** (JK ou JK-J × L × L) com o objetivo de determinar, usando ML-INDSCAL, a mesma matriz de pesos, **A**, diferentemente ponderada por cada fatia. Isto implica que todos os grupos compartilham um subespaço de estímulo comum. Por outro lado, se o tratamento $j = 1, \ldots, J$ representa diferentes intervenções (e.g., administração de uma droga A, B e placebo ou a investigação de indução de proteínas nativas e cepas com mutação, etc.), talvez não haja razão para acreditar que os tratamentos compartilham a mesma matriz de pesos, assim pode-se obter um modelo ML-INDSCAL para cada tratamento. Porém, a indicação do tipo de arranjo que deve ser usado deve ser dada, em último caso, pela questão biológica que fundamenta o experimento.

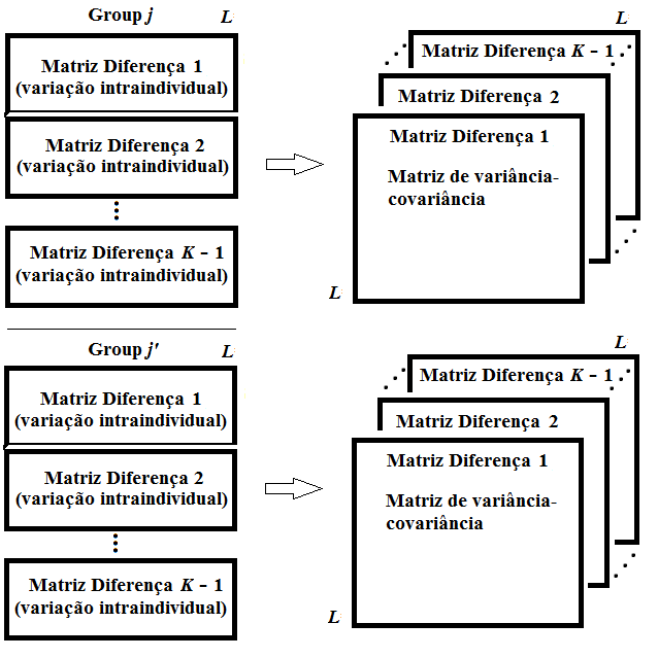

Figura 65: Montagem do arranjo para análise por ML-INDSCAL, por exemplo, a partir da separação da variância como na Figura 64. K é o número total de ocasiões (pontos de tempo) no estudo. L é o número de metabólitos (variáveis) no conjunto de dados. Os grupos de controle e tratamento são representados por $j = 1, \ldots, J$.

Contrastando o método ML-INDSCAL com os métodos baseados em PCA, o primeiro descreve as diferenças diretamente ao nível dos grupos experimentais, enquanto os outros acessam a distribuição dos indivíduos dentro dos grupos experimentais. Apesar disso, para o método ML-INDSCAL, uma visão aproximada ao nível dos indivíduos pode ser realizada investigando as variáveis apontadas como discriminatórias para os grupos experimentais, no dado original. Neste caso, as diferentes respostas dos indivíduos podem ser identificadas, além daqueles indivíduos mais sensíveis ao tratamento (com maior aumento ou diminuição nos níveis das variáveis). Da mesma forma, para os métodos baseados em PCA os níveis dos grupos podem ser examinados calculando escores médios [135].

A visualização do modelo ML-INDSCAL é similar aos outros modelos de componentes (como PCA ou PARAFAC [10]). As ponderações na matriz diagonal $\mathbf{G}_k$ ($R \times R$) indicam para a ocasião k quais as relações na matriz de pesos **A** são importantes, considerando cada componente. Por sua vez, a matriz de pesos **A** pode ser visualizada como gráficos de diferentes combinações de componentes, com o objetivo de verificar as variáveis relevantes (metabólitos, proteínas, genes, etc.) ou como foi proposto por Jansen *et al.* [113], através de gráficos de calor (*heat plots*) obtidos calculando o produto externo $\mathbf{a}_r\mathbf{a}_r^{\mathrm{T}}$, onde $\mathbf{a}_r$ é um vetor de pesos (uma coluna de **A**) relativo à componente principal R. A imagem simétrica no gráfico de calor segue a estrutura da matriz de covariância com dimensões ($L \times L$), então as importantes BVRs podem ser identificadas mais facilmente.

Outro ponto importante a ser considerado sobre o modelo ML-INDSCAL é o número de componentes, onde a escolha pode ser feita observando a variância explicada e o diagnóstico de consistência do núcleo (CORCONDIA) [11] que indica se o modelo determinado desvia da trilinearidade, pressuposta nos métodos ML-INDSCAL e PARAFAC. Geralmente, o modelo estará mais próximo da trilinearidade quanto maior for o valor de CORCONDIA, mas, uma vez que não há um valor

mínimo aceitável, é importante verificar a quantidade de variância explicada adicionada pelos componentes. A variância explicada somente pode ser calculada para o modelo inteiro sem partição entre as componentes, uma vez que os componentes no ML-INDSCAL (a exemplo do PARAFAC) não são ortogonais.

4.4.5. Validação usando a abordagem *jack-knife*

O procedimento *jack-knife* [12], proposto para calcular intervalos de confiança em modelos PARAFAC, consiste em um abordagem geral usada para situações onde aparentemente nenhum método melhor se encontra disponível. Os intervalos de confiança são calculados através de submodelos que são obtidos excluindo um número definido de amostras (uma amostra por ocasião neste trabalho) do conjunto de dados original. As amostras são consideradas independentes e pertencentes a uma mesma população. Para um arranjo de três modos $\underline{\mathbf{X}}$ ($I \times J \times K$) com I amostras, I submodelos PARAFAC são calculados deixando uma fatia (amostra) fora cada vez. O erro padrão pode ser determinado usando o conjunto de I valores de alguma estatística computada de cada submodelo. Na validação *jack-knife* dos modelos ML-INDSCAL, as estatísticas computadas são as ponderações na matriz $\mathbf{G}_k$, correspondendo a um componente em cada grupo experimental e a matriz de pesos $\mathbf{A}$, correspondendo a um componente em cada variável.

Diferentemente da abordagem *jack-knife* aplicada aos modelos PARAFAC, em ML-INDSCAL as amostras (indivíduos) não significam fatias no arranjo. Quando uma amostra por ocasião é deixada fora, as outras são usadas para computar as fatias, assim cada uma delas representa uma matriz de covariância com dimensões ($L \times L$) independente do número de amostras, uma vez que o número de variáveis são os mesmos para todas as amostras. Assim, os submodelos obtidos no método ML-INDSCAL sempre possuem arranjos com as mesmas dimensões do arranjo para o

modelo global. Em contraste, por exemplo, para aplicação em modelos PARAFAC de dados de fluorescência, cada fatia corresponde a uma matriz de emissão excitação, ou seja, uma amostra, e como resultado do procedimento *jack-knife* excluindo uma amostra por vez, há a diminuição de uma fatia por submodelo, levando a arranjos com uma amostra a menos do que o arranjo para o modelo global.

Sabe-se que os modelos PARAFAC sofrem de violações de unicidade [12], ou seja, os componentes do modelo podem ser permutados sem perda de ajuste e, além disso, eles possuem uma indeterminação de escalamento intrínseca. Para o modelo ML-INDSCAL, somente a indeterminação de permutação é importante, uma vez que devido às restrições impostas, o problema do escalamento é irrelevante [113]. Outro ponto importante a ser considerado é a intrínseca ambiguidade de sinal, que é similar àquela conhecida para a SVD ou decomposição em autovalores/autovetores [136]. A Equação 45 mostra um rearranjo do modelo usado em ML-INDSCAL, sendo os escalares g_{kr} os elementos na diagonal da matriz $\mathbf{G}_k$ relativa ao grupo experimental k no r-ésimo componente. O vetor de pesos do r-ésimo componente é expresso por $\mathbf{a}_r$. Nota-se que os produtos externos na Equação 45 produzem as mesmas matrizes simétricas, independente do sinal do vetor. Estas matrizes, conforme já citado, podem ser visualizadas como *heat plots*, não sendo as interpretações prejudicadas pela ambiguidade de sinal. Por outro lado, no contexto da abordagem *jack-knife* a indeterminação do sinal leva a problemas para a comparação de todos os submodelos, devido às diferentes orientações de cada vetor de pesos.

$$\mathbf{R}_k = \sum_{r=1}^{R} g_{kr}\mathbf{a}_r\mathbf{a}_r{}^{\mathrm{T}} + \mathbf{E} = g_{kr}(-\mathbf{a}_r)(-\mathbf{a}_r)^{\mathrm{T}} + \mathbf{E} \qquad \text{Equação 45}$$

Portanto, estas questões devem ser resolvidas antes da comparação direta dos submodelos obtidos. A solução para a indeterminação da permutação é obtida usando

a proposição de Riu & Bro [12], onde é procurada a ordem dos componentes que fornecem o coeficiente de congruência máximo entre os pesos de cada submodelo e os pesos do modelo global. Descrição detalhada para esta correção pode ser encontrada na literatura [12]. A ambiguidade de sinal é corrigida para todos os submodelos utilizando como padrão a orientação dada pelos sinais obtidos para os pesos do modelo global. Neste trabalho, uma rotina em código Matlab foi construída para a automatização destas correções.

Similarmente ao que foi proposto na literatura [113], 95% dos submodelos corrigidos que convergem para soluções estáveis e que possuem os menores desvios são usados para acessar as variabilidades dos valores estatísticos (valores de ponderações e de pesos). Os desvios dos submodelos são calculados a partir das suas ponderações, relacionando-as às ponderações do modelo global. Ao final, estas variabilidades podem ser visualizadas através de um gráfico dos valores estatísticos calculados para os submodelos juntamente com as estatísticas calculadas para o modelo global.

Neste trabalho, gráficos de caixas (*box plots*) são usados para visualizar também as diferenças entre as estatísticas dos submodelos, explorando a habilidade destes gráficos para mostrar a distribuição de valores, localização, assimetria, espalhamento, comprimento de "cauda", além de *outliers*. Portanto, as ocorrências de sobreposições das caixas ou das medianas, são observadas, com o intuito de verificar as diferenças entre os grupos de submodelos. As medianas são preferidas em relação às médias, pois *a priori* não há nenhuma suposição sobre a distribuição simétrica e/ou normal dos valores estatísticos.

4.5. Parte Experimental

4.5.1. Conjunto de dados simulado

Com o objetivo de ilustrar o método ML-INDSCAL e suas propriedades, um pequeno conjunto de dados simulado foi proposto (Tabela 6). O conjunto é composto por respostas medidas em quatro variáveis A, B, C e D (por exemplo, concentração de metabólitos ou medidas de expressão genética) em 10 indivíduos e coletadas em um período de tratamento, incluindo um controle, num planejamento experimental cruzado. A separação da variação seguiu o procedimento da Figura 63, ou seja, a matriz de dados, **C** (10 × 4), simulada como sendo relativa às respostas no período de controle, foi subtraída das matrizes, **T1** (10 × 4) e **T2** (10 × 4), referentes às ocasiões T1 e T2 no período de tratamento, fornecendo as matrizes $\mathbf{D}_{T1}$ (10 × 4) e $\mathbf{D}_{T2}$ (10 × 4) mostradas na Tabela 7, onde também podem ser vistas as matrizes de variância-covariância $\mathbf{R}_{DT1}$ (4 × 4) e $\mathbf{R}_{DT2}$ (4 × 4). Na Tabela 7, as covariâncias e variâncias podem ser visualizadas escolhendo-se apropriadamente uma linha e uma coluna nas matrizes $\mathbf{R}_{DT1}$ e $\mathbf{R}_{DT2}$, por exemplo, a covariância entre as variáveis A e C relativa à ocasião T1 igual a -10,56 pode ser vista na coordenada da primeira linha com a terceira coluna e o mesmo valor é observado na coordenada da terceira linha com a primeira coluna, dada a simetria da matriz. As matrizes $\mathbf{R}_{DT1}$ e $\mathbf{R}_{DT2}$ foram usadas para compor o arranjo $\underline{\mathbf{R}}$ (2 × 4 × 4), e então o método ML-INDSCAL foi aplicado. Os resultados são comparados aos obtidos com o método INDSCAL. Neste último caso, o arranjo é composto pelas matrizes de covariância de cada matriz **C**, **T1** e **T2** e tem dimensão (3 × 4 × 4). Além disso, o método PCA foi aplicado na matriz obtida centrando os dados de **T** na média, que por sua vez foi obtida pela concatenação das matrizes **C**, **T1** e **T2** e na matriz obtida centrando os elementos de **D** na média, construída pela concatenação das matrizes $\mathbf{D}_{T1}$ e $\mathbf{D}_{T2}$ (este último é equivalente ao

método PRC). Antes de todas as análises, uma pequena fração de ruído aleatório e normalmente distribuído foi adicionado aos dados.

4.5.2. Conjunto de dados real: exemplo de expressão genética

Este conjunto de dados é de um estudo de perfis de expressão celular global, onde linhas de células expressando a proteína Vpr na forma nativa e com mutações (cepas R80A e F72A/R73A) são investigadas [119]. Após indução com doxiciclina, RNA foi coletado nos tempos de 0, 1, 2, 4, 6, 8, 12, 16 e 24 h. O experimento foi realizado com três replicatas biológicas a cada ocasião. Expressão genética foi medida usando microarranjos de oligonucleotídeos contendo cerca de 22.400 elementos em cada arranjo.

Para avaliação da expressão diferencial, os perfis das células nativas e mutantes foram comparados por testes de permutação multivariados baseados num teste *t* estatístico para uma amostra com variância aleatória. Além disso, genes celulares representando um efeito biológico plausível foram arbitrariamente considerados com mudanças de pelo menos duas vezes na expressão. Com base nisto e comparando os perfis de expressão das linhas celulares expressando a proteína Vpr nativa e a mutante F72A/R73A, 121 genes significativos foram encontrados, mas entre estes apenas 47 foram eventos específicos da forma nativa. Quando a comparação foi entre as linhas celulares expressando a proteína Vpr nativa e a mutante R80A, 66 genes significativos foram encontrados sendo que 42 deles como eventos específicos da forma nativa. Os dados relativos aos genes manifestados especificamente para a forma nativa foram estudados neste trabalho por ML-INDSCAL e PCA, sendo esta última realizada sobre os dados centrados na média. Todos os dados são disponíveis publicamente [119,120] e descrições detalhadas das estratégias estatísticas comentadas podem ser encontradas na literatura [119].

Tabela 6: Dados simulados com respostas de quatro variáveis (A, B, C e D) em 10 indivíduos coletados em dois períodos T1 e T2, e um período controle.

Tratamento	Indivíduos	A	B	C	D
Controle (C)	1	25	123	26	58
	2	98	62	60	22
	3	17	77	86	72
	4	16	104	118	32
	5	60	108	66	63
	6	80	96	36	28
	7	41	56	76	77
	8	120	121	113	57
	9	68	80	95	111
	10	46	112	46	45
T1	1	68,9	139,1	78,8	132
	2	142,9	78,1	111,8	96,5
	3	62,9	93,1	136,8	146
	4	62,9	120,1	167,8	106,5
	5	107,9	124,1	114,8	137
	6	128,9	112,1	82,8	102,5
	7	90,9	72,1	121,8	151
	8	170,9	137,1	157,8	131,5
	9	119,9	96,1	138,8	185
	10	98,9	128,1	88,8	119,5
T2	1	69,4	139,1	68,8	132,5
	2	143,4	78,1	103,8	97,5
	3	63,4	93,1	130,8	146,5
	4	63,4	120,1	163,8	107,5
	5	108,4	124,1	112,8	137,5
	6	129,4	112,1	84,8	103,5
	7	91,4	72,1	125,8	151,5
	8	171,4	137,1	163,8	132,5
	9	120,4	96,1	146,8	185,5
	10	99,4	128,1	98,8	120,5

Tabela 7: Matrizes de diferenças relativas ao conjunto de dados simulado na Tabela 6; $\mathbf{D}_{T1}$ e $\mathbf{D}_{T2}$ correspondem aos dois períodos de tratamento, T1 e T2; As matrizes $\mathbf{R}_{DT1}$ e $\mathbf{R}_{DT2}$ referem-se às matrizes de covariância destes mesmos períodos.

Matrizes	Indivíduos/variáveis	A	B	C	D
$\mathbf{D}_{T1}$	1	43,9	16,1	52,8	74
	2	44,9	16,1	51,8	74,5
	3	45,9	16,1	50,8	74
	4	46,9	16,1	49,8	74,5
	5	47,9	16,1	48,8	74
	6	48,9	16,1	46,8	74,5
	7	49,9	16,1	45,8	74
	8	50,9	16,1	44,8	74,5
	9	51,9	16,1	43,8	74
	10	52,9	16,1	42,8	74,5
$\mathbf{D}_{T2}$	1	44,4	16,1	42,8	74,5
	2	45,4	16,1	43,8	75,5
	3	46,4	16,1	44,8	74,5
	4	47,4	16,1	45,8	75,5
	5	48,4	16,1	46,8	74,5
	6	49,4	16,1	48,8	75,5
	7	50,4	16,1	49,8	74,5
	8	51,4	16,1	50,8	75,5
	9	52,4	16,1	51,8	74,5
	10	53,4	16,1	52,8	75,5
$\mathbf{R}_{DT1}$	A	9,17	0	-10,56	0,14
	B	0	0	0	0
	C	-10,56	0	12,22	-0,17
	D	0,14	0	-0,17	0,07
$\mathbf{R}_{DT2}$	A	9,17	0	10,56	0,28
	B	0	0	0	0
	C	10,56	0	12,22	0,33
	D	0,28	0	0,33	0,28

4.6. Resultados e Discussão

4.6.1. Conjunto de dados simulado

4.6.1.1. Método ML-INDSCAL

A Figura 66 mostra os resultados da abordagem ML-INDSCAL sobre o conjunto de dados simulado. Dois componentes foram escolhidos para o modelo observando a variância explicada de 99% e uma CORCONDIA de 98%. A Figura 66A ilustra as ponderações das "fatias" compostas pelas matrizes de covariância, $\mathbf{R}_{DT1}$ e $\mathbf{R}_{DT2}$, nos componentes do modelo ML-INDSCAL. A ponderação do grupo associado à ocasião T1 é igual a 20,44 na componente 1, enquanto a ponderação do grupo associado à ocasião T2 tem valor 0,26. Por outro lado, na componente 2, a ponderação do primeiro é 0,30 e a do último é 19,44. Uma vez que os grupos são tomados em relação ao grupo controle, o fato dos mesmos possuírem ponderações não nulas permite concluir que ML-INDSCAL é capaz de distingui-los do grupo de referência. Além disso, como pode ser visto na Figura 66A, as ponderações em ML-INDSCAL indicam uma clara discriminação entre os grupos experimentais (amostras em diferentes ocasiões T1 e T2) e as ponderações para os submodelos (pontos coloridos na Figura 66A), obtidos através da validação *jack-knife*, apontam a significância estatística desta diferença.

Os pesos do modelo ML-INDSCAL são mostrados na Figura 66B, enquanto as Figuras 66C e 66D apresentam os *heat plots* dos componentes 1 e 2. Nota-se, a partir dos pesos na Figura 66B que as variáveis A e C são as mais importantes para a discriminação entre as ocasiões, o que significa que ou suas variâncias ou covariâncias estão mudando e, consequentemente, levando à distinção durante o

tempo considerado. Neste caso, o resultado surge devido às covariâncias com a inversão nos seus sinais, uma vez que as variâncias das variáveis A e C permanecem as mesmas em T1 e T2 (veja Tabela 7). A afirmação é corroborada pelos *heat plots* na Figura 66C, onde pode ser visto um quadrado azul (nas coordenadas (A,C) e (C,A)) representando um valor negativo, enquanto na mesma posição no *heat plot* da Figura 66D, o quadrado representa um valor positivo.

Embora a covariância entre as variáveis C e D apresentem o mesmo comportamento (mudança de sinal), elas não são capturadas pelas componentes, do mesmo modo que a variância da variável D que possui um aumento de quatro vezes na magnitude. Estas partes do dado contam com menores valores nas "fatias" e por esta razão, provavelmente, não são envolvidas na discriminação. Vale ressaltar que está é uma situação hipotética, diretamente associada às variâncias e covariâncias simuladas. Para uma análise real, os resultados irão surgir a partir da variância em questão, que depende das manipulações experimentais envolvidas e de BVRs específicas. O dado simulado mostra um exemplo extremo para mudança na covariância, porém situações intermediárias podem ser encontradas em sistemas biológicos reais.

As Figuras 66E e 66F contrastam as dispersões dos dados originais para as variáveis A e C, após e antes da separação na variância seguindo a proposição em ML-INDSCAL, sendo o gráfico na Figura 66E muito similar ao paradigma mostrado na Figura 61C. Na Figura 66F o mesmo paradigma permanece não revelado devido à abundancia absoluta das variáveis, que estão também sendo levados em conta junto com o efeito líquido.

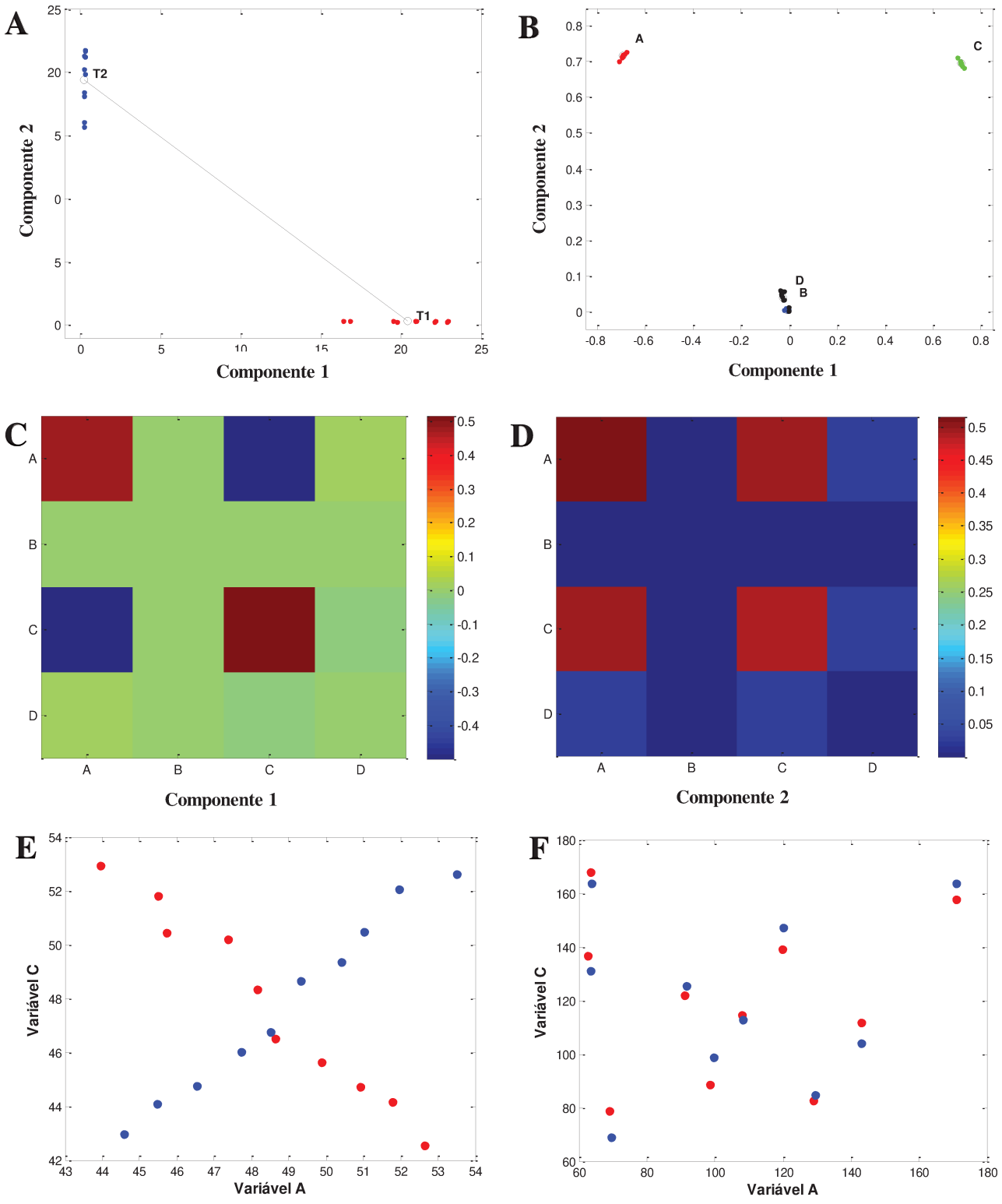

Figura 66: Modelo ML-INDSCAL do exemplo simulado. A: Ponderações dos grupos em T1 e T2 para componente 1 *versus* componente 2 e para os submodelos *jack-knife* (pontos coloridos); B: Pesos das variáveis A, B, C e D para a componente 1 *versus* a

componente 2 e para os submodelos *jack-knife* (pontos coloridos); C: Pesos para a componente 1 apresentados com *heat plot*; D: Pesos para a componente 2 apresentados com *heat plot*; E: Variável A *versus* variável C das matrizes D_{T1} (pontos vermelhos) e D_{T2} (pontos azuis); F: Variável A *versus* variável C das matrizes T1 (pontos vermelhos) e T2 (pontos azuis).

Métodos como PCA e INDSCAL não focam somente no efeito líquido e não removem a magnitude grosseira dos dados (removendo a pré-dose ou a resposta do grupo controle). Com o objetivo de comparar com os resultados do modelo ML-INDSCAL, um modelo INDSCAL com quatro componentes (ambas, variância explicada e CORCONDIA, cerca de 99%) foi construído usando um arranjo **R** (3 × 4 × 4) composto pelas matrizes de covariância das matrizes **C**, **T1** e **T2**, e uma modelo PCA da matriz **T** obtida pela concatenação vertical de **C**, **T1** e **T2** (Tabela 6) foi também construído.

4.6.1.2. ML-INDSCAL *versus* INDSCAL

As Figuras 67A e 67B mostram os resultados do modelo INDSCAL sobre os dados simulados. As ponderações de cada grupo experimental nas três primeiras componentes não apresentam diferenças significativas, como observado através das ponderações dos submodelos (pontos coloridos) obtidos com o procedimento *jack-knife*, na Figura 67A. Nota-se apenas uma pequena diferença nas ponderações para o grupo T1, na quarta componente, mas como pode ser visto na Figura 67B, os pesos nesta componente não atingem diferenças significativas.

Os resultados citados acima podem estar associados à presença da variação não dinâmica contida nos dados, que dificultam a evidência do efeito líquido discriminatório. A Tabela 8 mostra as matrizes de covariância para os grupos experimentais sem a separação na variância, onde é possível contrastar seus altos valores, devido à variação não dinâmica, com os valores nas matrizes de covariância

na Tabela 7. Os *heat plots* relativos aos quatro componentes são mostrados na Figura 68 e indicam BVRs mais complexas. Portanto, INDSCAL não diferencia os grupos T1 e T2 entre si e nem estes em relação ao grupo controle.

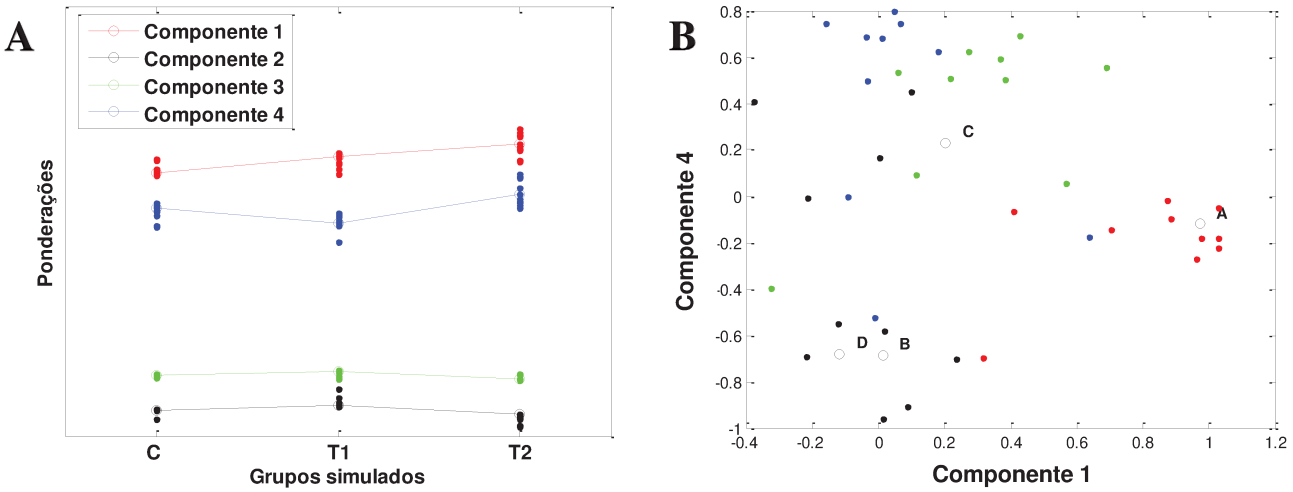

Figura 67: Modelo INDSCAL do exemplo simulado. A: ponderações dos grupos **C**, **T1** e **T2** para as quatro componentes e para os submodelos *jack-knife* (pontos coloridos); B: pesos para as variáveis A, B, C e D (componente 1 *versus* componente 4) e para os submodelos *jack-knife* (pontos coloridos).

Embora a variável B não mostre qualquer modificação ao longo das ocasiões simuladas T1 e T2, há BVRs incluindo-a, provavelmente devido ao seu nível inicial no grupo controle. De fato, há um aumento no nível da variável B a partir do grupo controle até o grupo T1 e que permanece constante. Esta modificação não é contemplada pelo ML-INDSCAL, pois, este remove a informação do grupo controle com o objetivo de focar apenas nas mudanças ao longo do tempo do tratamento. As mudanças capturadas por ML-INDSCAL seguem os paradigmas multivariados, que não é caso da variável B, uma vez que ela não tem qualquer correlação significante com as demais variáveis. Para algumas situações como esta, um método univariado pode ser usado para auxiliar na investigação. Tais métodos estão fora do escopo deste trabalho.

Tabela 8: Matrizes de covariância dos grupos experimentais no conjunto de dados simulado sem a separação na variância proposta em ML-INDSCAL.

Matrizes	Variáveis	A	B	C	D
$\mathbf{R}_C$*	A	1,22	0,01	0,09	-0,14
	B	0,01	0,57	-0,08	-0,12
	C	0,09	-0,08	0,97	0,23
	D	-0,14	-0,12	0,23	0,71
$\mathbf{R}_{T1}$*	A	1,29	0,01	0,06	-0,11
	B	0,01	0,57	-0,08	-0,12
	C	0,06	-0,08	0,95	0,20
	D	-0,11	-0,12	0,20	0,70
$\mathbf{R}_{T2}$*	A	1,28	0,01	0,15	-0,11
	B	0,01	0,57	-0,07	-0,12
	C	0,15	-0,07	1,03	0,27
	D	-0,11	-0,12	0,27	0,69

* Os valores encontram-se divididos por 10^3; Os subscritos indicam o grupo referente.

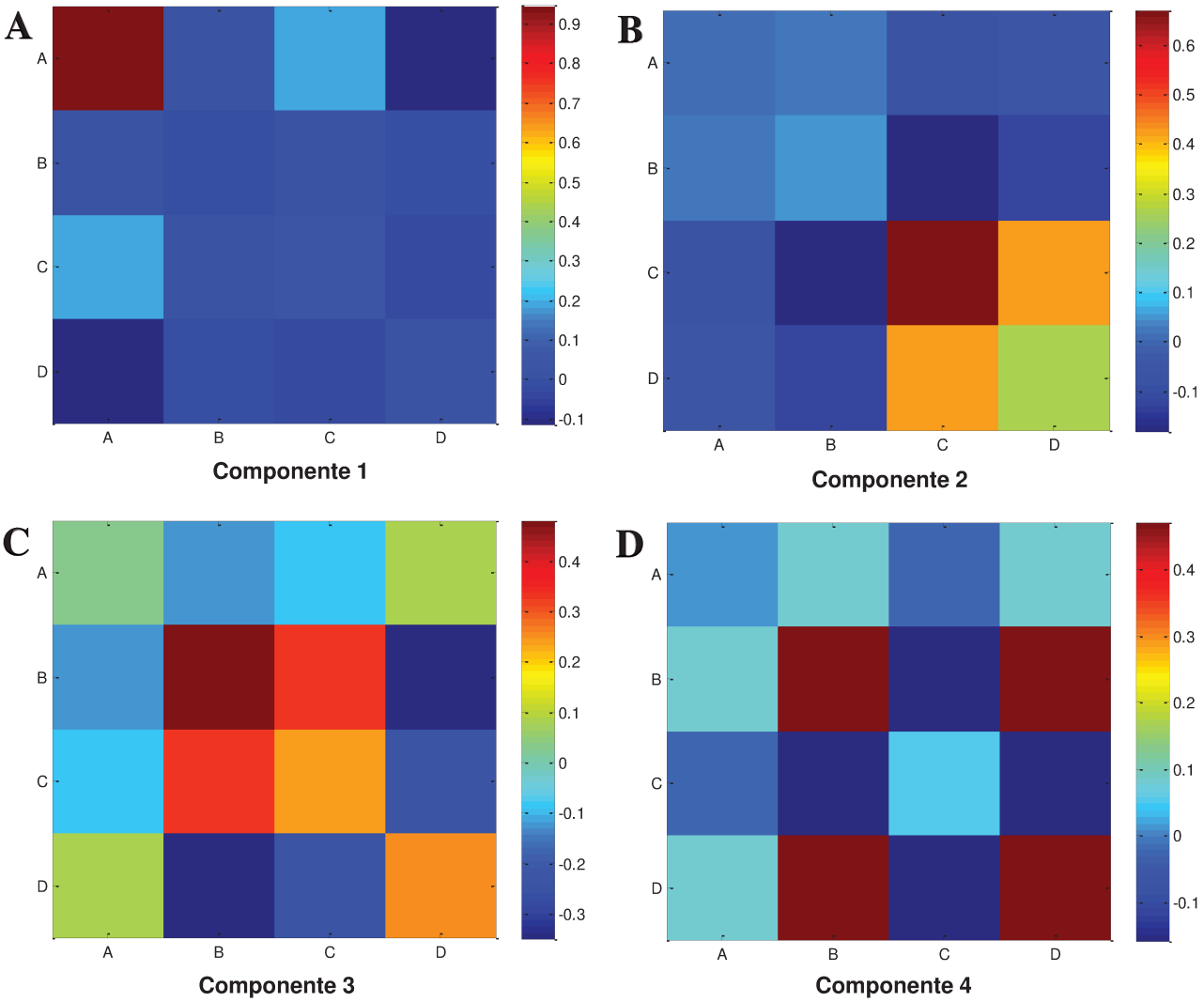

Figura 68: Gráficos de calor (*heat plots*) relativos às quatro componentes principais obtidas no modelo INDSCAL para o conjunto de dados simulado.

4.6.1.3. ML-INDSCAL *versus* PCA

Os resultados da PCA aplicada aos dados simulados são apresentados na Figura 69. Como observado no gráfico de escores da Figura 69A, a variância capturada em PC1 revela que parte das respostas simuladas do tratamento corresponde a uma relação positiva entre os níveis das variáveis A, C e D. Relacionado a esta resposta em PC1, percebe-se que é possível distinguir somente os indivíduos no grupo controle

daqueles nos tempos, T1 e T2, não se observando diferenças entre estes dois últimos grupos. Isto indica que a separação é principalmente devido à abundância absoluta das variáveis (variação não dinâmica) que diferentemente do ML-INDSCAL, em PCA não é removida.

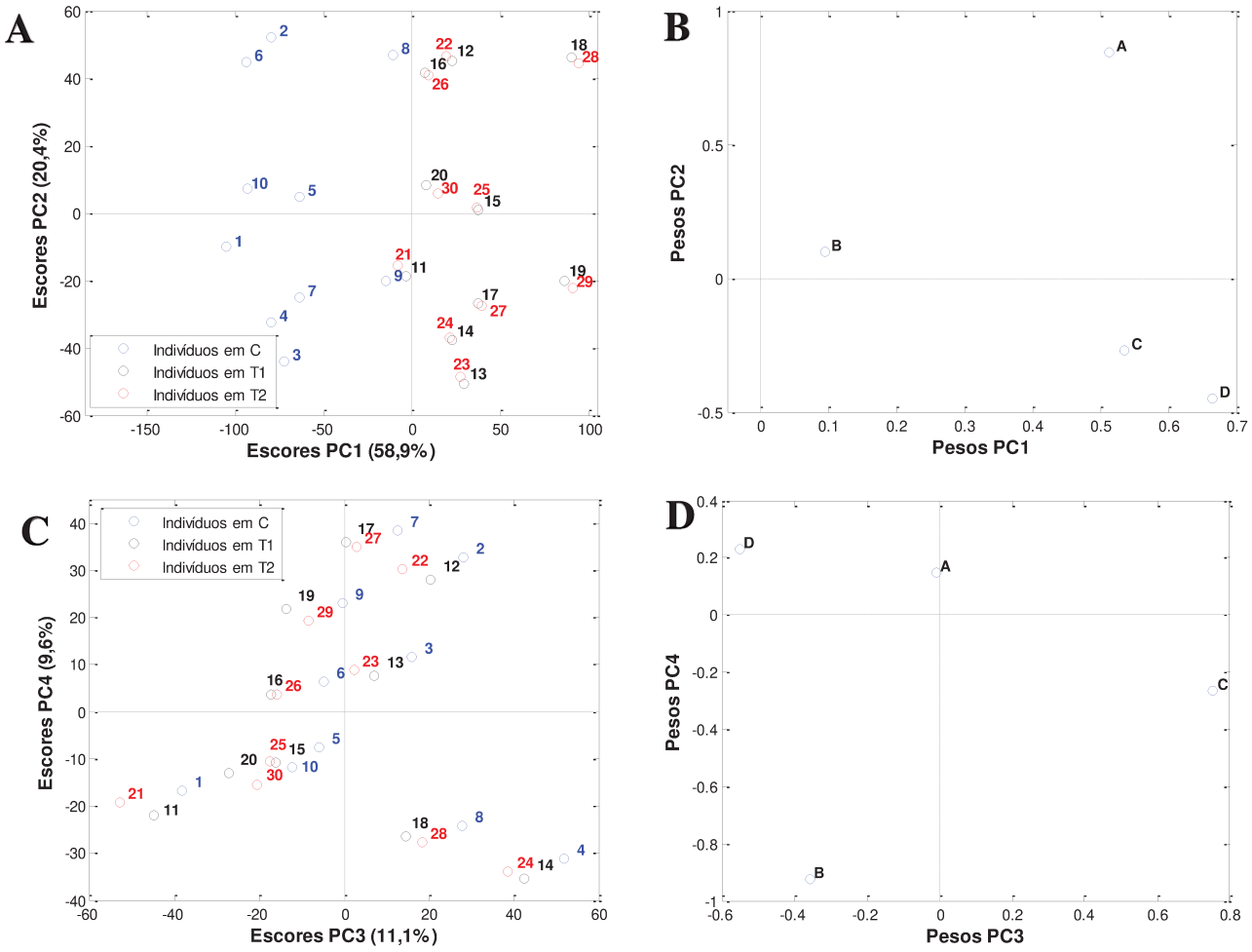

Figura 69: Modelo PCA para os dados simulados. A: Gráfico de escores PC1 *versus* PC2; B: Gráfico de pesos PC1 *versus* PC2; C: Gráfico de escores PC3 *versus* PC4; D: Gráfico de pesos PC3 *versus* PC4.

Outra parte da resposta é a variância ortogonal capturada em PC2 mostrando uma relação negativa entre a variável A e as variáveis C e D, basicamente indicando os indivíduos que possuem, simultaneamente, níveis mais altos da primeira e mais baixos níveis das últimas, com escores positivos, além daqueles com característica

oposta, com escores negativos. Esta relação não é necessariamente ao longo do tempo de tratamento, sendo notada até mesmo no grupo controle. A variável B apresenta pouca ou nenhuma correlação com as demais variáveis e assim, mostra pesos próximos de zero para ambas as componentes principais.

Observando o gráfico de escores para a terceira e quarta componente principal na Figura 69C, não é observada qualquer separação entre os grupos, estando as variáveis descrevendo características específicas de alguns indivíduos, ou seja, possivelmente a variação interindividual, não interessante no contexto do tratamento.

4.6.1.4. ML-INDSCAL *versus* PRC

Um modelo PRC foi construído aplicando PCA sobre a matriz **D** contendo apenas a variação intra-individual, obtida pela concatenação vertical das matrizes $\mathbf{D_{T1}}$ e $\mathbf{D_{T2}}$ (Tabela 7). Este modelo é apresentado na Figura 70. As duas primeiras componentes principais explicam cerca de 91% da variância intra-individual e revelam que a resposta ao tratamento é basicamente devido às variáveis A, em PC2 e C, em PC1.

O gráfico de escores na Figura 70A apresenta a distribuição dos indivíduos que é similar àquela obtida para os dados originais (Figura 66E) correspondendo aos indivíduos com mais altos níveis da variável C, com escores negativos em PC1 e mais altos níveis da variável A, com escores negativos em PC2. Apesar disso, o modelo não foi capaz de evidenciar diferenças entre os grupos experimentais, como mostrou o ML-INDSCAL "explorando" a mudança na covariância entre as variáveis A e C. Como pode ser visto na Figura 70B, estas variáveis respondem sozinhas em cada componente principal, assim não apresentando correlação, uma vez que as componentes são ortogonais. De fato, PRC utilizando a PCA extrai as componentes computando uma matriz de covariância incluindo todos os indivíduos (ou grupos

experimentais), sendo a informação sobre as mudanças nas covariâncias (ou correlações) perdidas. A Tabela 9 mostra a matriz de covariância citada, onde é possível ver o pequeno valor da covariância entre as variáveis A e C, em relação às variâncias destas.

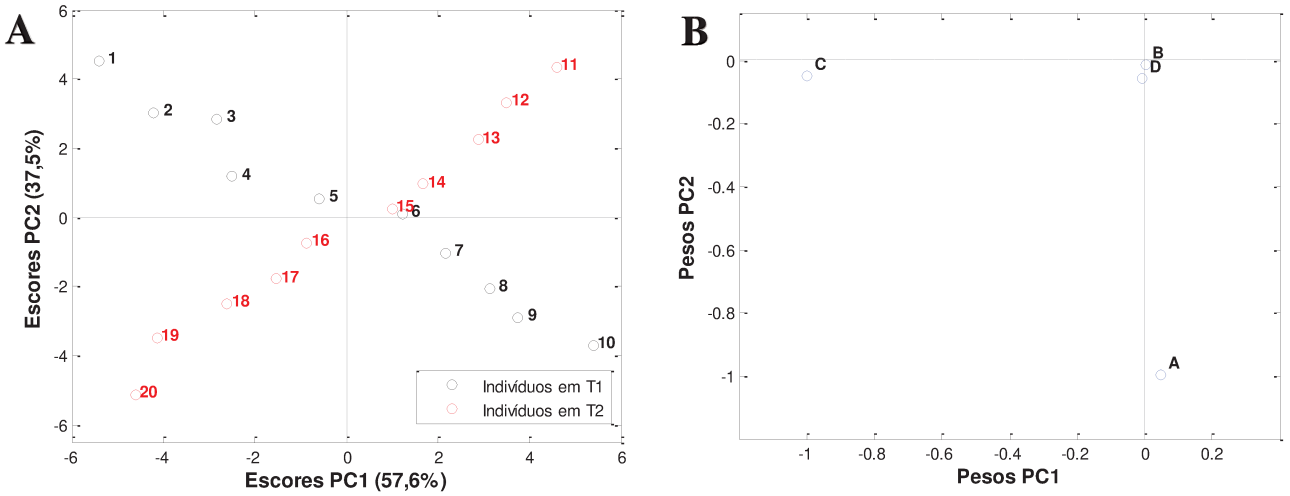

Figura 70: Modelo PRC para os dados simulados. A: Gráfico de escores PC1 *versus* PC2; B: Gráfico de pesos PC1 *versus* PC2.

O exemplo simulado mostra o método ML-INDSCAL como um procedimento capaz de trazer informações sobre as variáveis que seguem o paradigma da Figura 61C, assim fornecendo uma ferramenta útil para a seleção de características quando a questão biológica que fundamenta o tratamento ou intervenção tem supostamente BVRs como esta, que é indetectável pelas análises padrões de componentes. Não somente o paradigma citado, que é uma situação extrema onde o sinal da covariância inverte, mas é esperado que a abordagem ML-INDSCAL também evidencie mudanças nas magnitudes seja das covariâncias ou variâncias. Como foi demonstrado, ML-INDSCAL diferenciou os grupos em T1 e T2 entre si e também em relação ao grupo controle, ao contrário dos demais métodos que na melhor situação apenas conseguem fazer a distinção entre a intervenção e o controle.

Tabela 9: Matriz de covariância para a matriz **D** obtida pela concatenação das matrizes $\mathbf{D_{T1}}$ e $\mathbf{D_{T2}}$.

Matriz	Variáveis	A	B	C	D
$\mathbf{R_D}$	A	7,22	0,21	-0,29	0,53
	B	0,21	0,19	-0,02	0,13
	C	-0,29	-0,02	11,13	0,08
	D	0,53	0,13	0,08	0,81

ML-INDSCAL pode ser usado como uma análise complementar revelando biomarcadores que proporcionam uma visão diferente sobre o sistema biológico. Por fim, uma vez que as interpretações sobre as BVRs a partir de matrizes de covariância podem ser tediosas e imprecisas para conjuntos de dados grandes (geralmente com centenas de variáveis), métodos como a ML-INDSCAL são altamente úteis e desejáveis, pois podem fornecer uma visão holística através de um modelo de componentes, focando na variação mais importante do conjunto de dados.

4.6.1. Conjunto de dados real: exemplo de expressão genética

Este conjunto de dados [119,120] é resultado de medidas simultâneas de muitos genes (cerca de 22.400) e diz respeito ao campo da virologia, especificamente no estudo da proteína viral R (Vpr, do inglês *Viral Protein R*) do vírus da imunodeficiência humana do tipo 1 (HIV-1, do inglês *Human Immunodeficiency Virus type* 1). O papel desta proteína na patogênese do HIV é investigado, observando-se o efeito da mesma na indução da interrupção do ciclo celular na fase G2/M [119]. A expressão genética foi medida a partir da proteína na sua forma nativa e em duas linhas com mutação, F72A/R73A e R80A, derivadas da forma nativa. As amostras foram medidas em nove pontos no tempo (0, 1, 2, 4, 6, 8, 12, 16 e 24 h). Neste exemplo, as amostras da forma nativa são consideradas como grupo de tratamento e as linhas com mutação são consideradas como grupos de controle, uma

vez que o efeito de interrupção do ciclo celular é observado apenas para as primeiras, sendo as últimas incapazes de produzi-lo. Deste modo, o conjunto de dados permite que dois modelos ML-INDSCAL, utilizando uma separação na variância similar ao método PRC, sejam construídos, um para cada grupo de controle. O conjunto de dados é baseado num planejamento experimental paralelo, pois diferentes indivíduos são medidos em cada tratamento e a cada tempo, portanto, as médias dos grupos de controle são utilizadas na subtração para separação da variância.

Para estes modelos, o procedimento *jack-knife* foi realizado com várias repetições, deixando fora de cada submodelo a primeira amostra de cada tempo, sendo as amostras nos grupos permutadas aleatoriamente de posição entre as construções dos submodelos. Isto foi realizado para acessar várias combinações de amostras que são deixadas fora de cada grupo, uma vez que os sujeitos são diferentes dentro dos mesmos. Ao final do procedimento, o número de submodelos é igual ao número de repetições. Vale ressaltar que se o planejamento experimental fosse cruzado, o número de submodelos seria igual ao número de amostras nos grupos, pois a aleatorização acima não seria necessária, tendo em vista que, neste caso, o mesmo "indivíduo" poderia ser deixado fora em cada grupo para a construção de cada submodelo.

O modelo ML-INDSCAL comparando o grupo de amostras nativas com o grupo de amostras com a mutação R72A/R73A (controle) utilizou os 47 genes observados como eventos específicos das primeiras. Assim, um modelo com duas componentes foi obtido com variância explicada de 77,2% e CORCONDIA de 90,6%. Para a validação deste modelo, 1000 repetições foram utilizadas para a construção de submodelos *jack-knife*, conforme descrito no item 4.4.5. Foi observado que modelos com um número maior de componentes não explicavam uma quantidade de variância estável, apresentando problemas de convergência para soluções estáveis, além de baixos valores de CORCONDIA (menores que 50%).

A Figura 71 apresenta os resultados deste modelo, onde na primeira componente nota-se a diferença das ponderações dos grupos em 4 e 16 h e na segunda componente a diferença da ponderação do grupo em 1 h (as ponderações citadas possuem os maiores valores). As significâncias destas diferenças são mostradas através dos gráficos de caixas (*box plots*) nas Figuras 71C e 71E, onde se observa para os grupos citados pouca ou nenhuma sobreposição para suas caixas e medianas. O modelo mostra que na resposta da expressão genética da Vpr são significamente importantes (veja os gráficos de caixas nas Figuras 71D e 71F) as variâncias e BVRs dos genes: TRIM65 (*tripartite motif-containing protein 65*); TARBP1 (*TAR (HIV) RNA binding protein 1*); HSPC023 (*HSPC023 protein - chromosome 19 open reading frame 53*); GMFB (*Glia Maturation Factor Beta*); MTMR1 (*Myotubularin Related Protein 1*); e AFAP (*Actin Filament-Associated Protein*). As relações entre os genes podem ser vistas diretamente na Figura 72, onde são mostrados os gráficos de calor obtidos a partir dos pesos do modelo ML-INDSCAL. A Tabela 10 apresenta uma breve descrição dos genes citados.

Os pesos das variáveis na primeira componente evidenciam principalmente os genes que são regulados para cima (*up-regulation* dos primeiros 19 genes), com destaque para o gene AFAP e os genes que são regulados para baixo (*down-regulation* a partir do gene na posição 20), sendo destacados na validação os genes GMFB, TRIM65, MTMR1 e TARBP1.

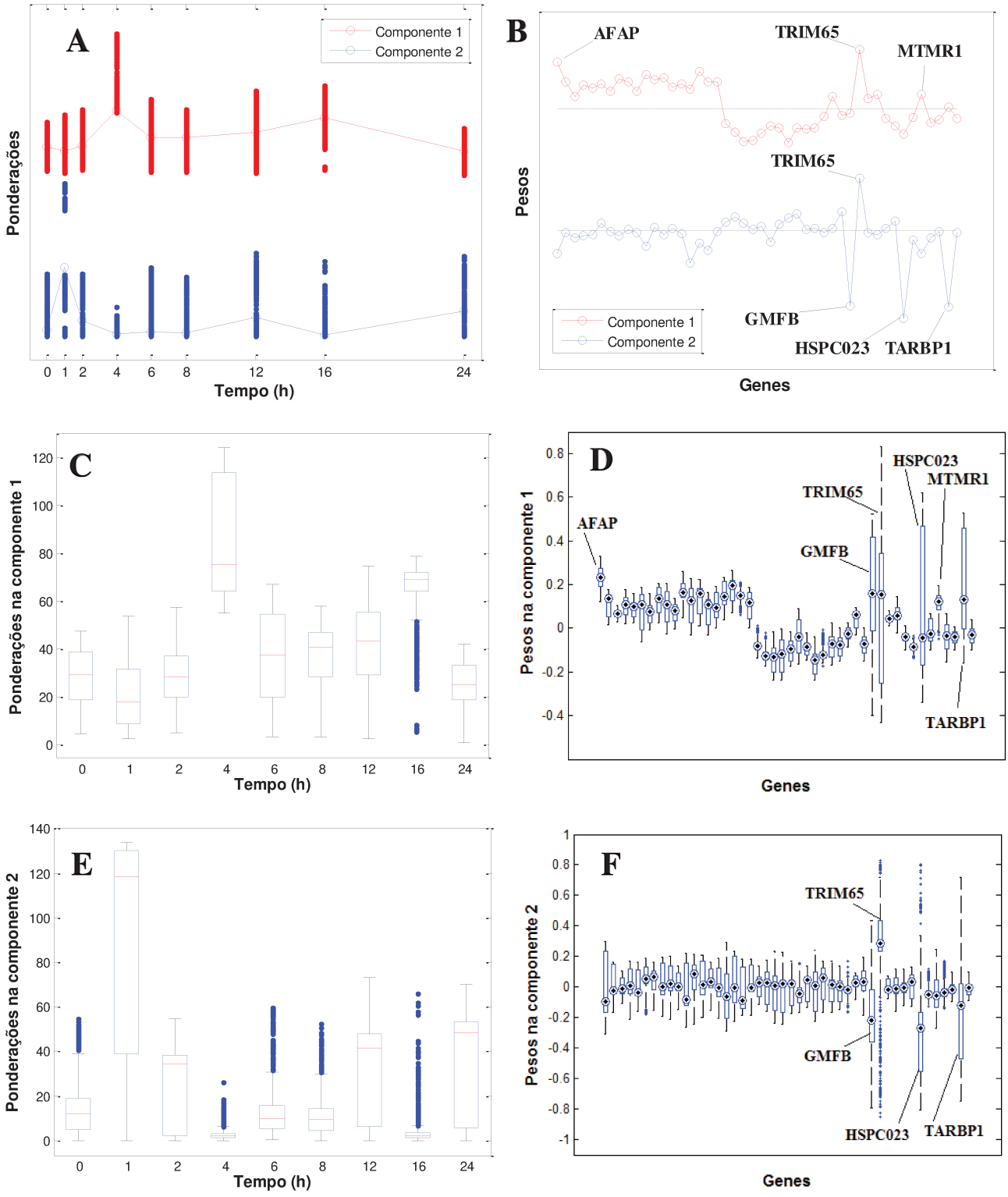

Figura 71: Modelo ML-INDSCAL com dois componentes para os 47 genes expressos especificamente para a Vpr na forma nativa comparando com a mutação R72A/F73A. A: Ponderações para os grupos experimentais e submodelos *jack-knife* (pontos coloridos); B: Pesos para as variáveis (genes), a linha cinza indica o zero; Gráficos de

caixa obtidos através das estatísticas dos submodelos *jack-knife*. C: Ponderações na componente 1; D: Pesos na componente 1; E: Ponderações na componente 2; F: Pesos na componente 2. Em A e B os valores estão deslocados verticalmente para uma melhor visualização e suas escalas são similares aos respectivos gráficos em caixas.

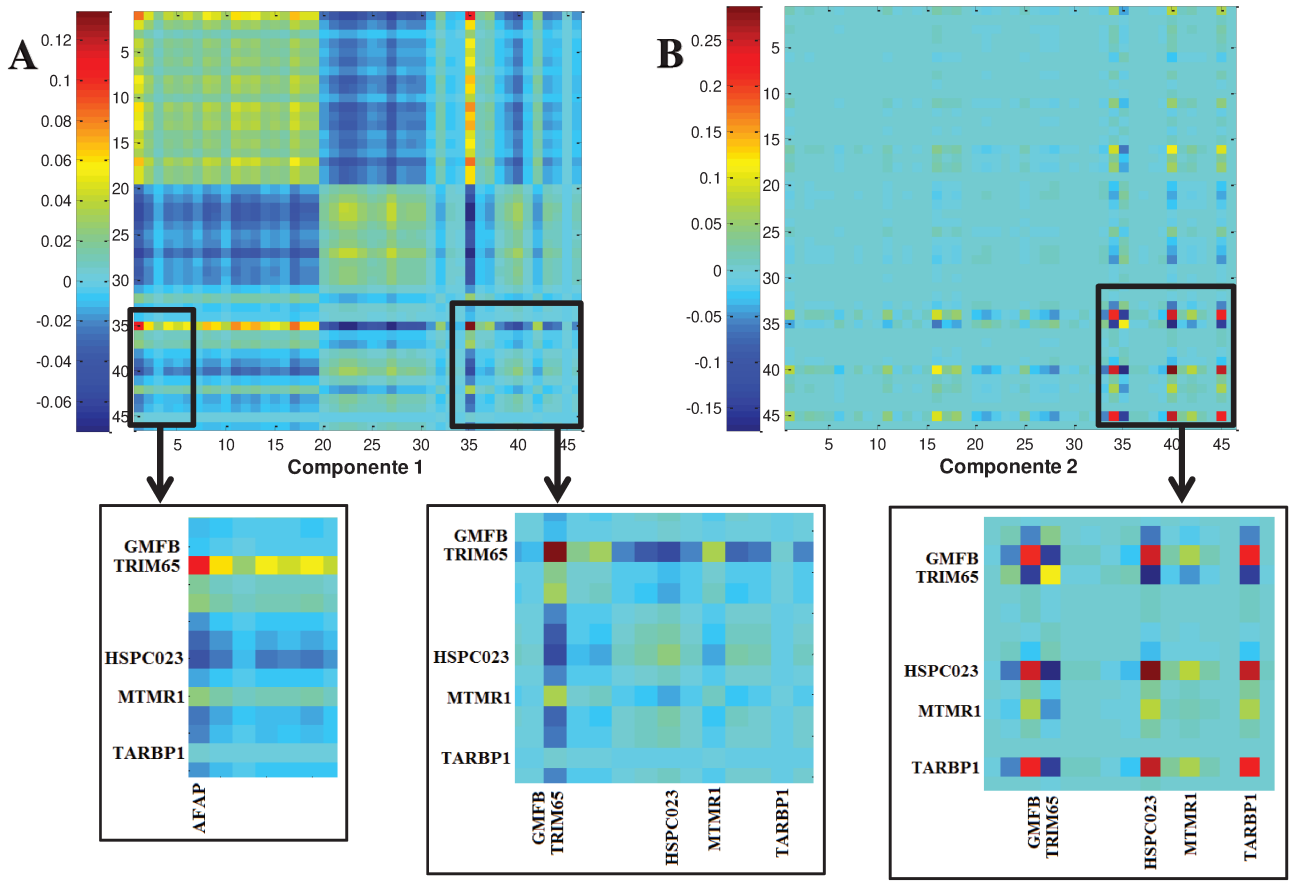

Figura 72: Gráficos de calor obtidos a partir dos pesos do modelo ML-INDSCAL na Figura 71. A: Componente 1; B: Componente 2. Em cada gráfico, regiões ampliadas com os genes mais importantes são mostradas.

O gene TARBP1 codifica a proteína TARBP1 envolvida na regulação transcricional, devido à sua habilidade de ligar-se a uma região em laço de um elemento regulatório de RNA (TAR) no HIV-1, uma região também ligada à RNA polimerase II. A proteína TARBP1 pode atuar separando a RNA polimerase II da TAR durante a elongação transcricional, sendo as ligações entre TARBP1 ou RNA

polimerase II à TAR, mutualmente exclusivas [137,138]. Já, o gene TRIM65 pode estar associado à inserção da sequência genética do HIV dentro do genoma do hospedeiro, uma vez que a proteína codificada por este gene pode ligar-se às cadeias repetidas de DNA do HIV (LTR, do inglês *Long Terminal Repeat*) [139].

A proteína GMFB expressa pelo gene GMFB inibe a atividade de proteínas quinases ativadas por mitógenos (MAPK, do inglês *Mitogen-Activated Protein Kinase*) em um ensaio *in vitro*, atuando em particular nas proteínas ERK1/ERK2 que possuem papel na progressão do ciclo celular e estão situadas, na rota de sinalização das MAPK, imediatamente acima das proteínas quinases citadas por Yoshizuka *et al.* [119] como importantes alvos da Vpr do vírus HIV para indução da interrupção do ciclo celular na fase G2/M [140,141]. Por outro lado, a proteína MTMR1 apresenta atividade catalítica da hidrólise de fosfatos ligados aos resíduos de tirosina em proteínas participando de diversos mecanismos biológicos. Finalmente, o gene AFAP está envolvido com as mudanças na integridade do citoesqueleto de actina que sofre grandes modificações e remodelagens durante a progressão celular [142].

Conforme informado pelo modelo na primeira componente, todos os genes citados acima têm relação positiva entre suas variâncias em 4 e 16 h, possivelmente indicando a transcrição do material genético do HIV, que por sua vez pode combinar a inserção da sua fita de DNA (TRIM65), o uso de estratégias transcricionais (TARBP1) e a inibição do ciclo celular do hospedeiro (GMFB) para imposição do seu próprio. Tais mecanismos podem ainda, valendo-se da maquinaria biosintética da célula hospedeira, promover o uso de proteínas auxiliares, além de proporcionar mudanças no citoesqueleto (AFAP), tendo em vista a alteração da progressão celular natural. Vale ressaltar que Yoshizuka *et al.* [119] mostrou que antes de 6 h, a Vpr do HIV-1 regula para baixo genes específicos que são relacionados à progressão do ciclo celular e assim alcançando a interrupção deste ciclo em cerca de 16 h. Portanto, os resultados aqui apontados suportam estas evidencias.

A segunda componente mostra relações negativas entre o gene TRIM65 (mais alta variância) e cada um dos genes GMFB, HSPC023 e TARBP1 ocorrendo no início da infecção, em 1 h. Pode-se hipotetizar que nesta ocasião as variâncias mais baixas destes últimos genes decorrem de uma estratégia da célula hospedeira como resposta imunológica à entrada do material genético viral, que conforme mostrado pela primeira componente, tende a sucumbir ante ao domínio do HIV, que alcança a otimização das suas funções nas ocasiões posteriores da infecção. O gene HSPC023 expressa uma proteína sem função ainda bem definida, sendo associada a um papel potencial na hipercalcemia da malignidade, e a partir do que se conhece até agora, ela nunca antes havia sido associada à infecção por HIV.

O modelo ML-INDSCAL comparando o grupo de amostras nativas com um grupo de controle (amostras com a mutação R80A) utilizou 42 genes especificamente expressos. Neste caso, um modelo com duas componentes foi obtido com variância explicada de 89,9% e CORCONDIA de 97,5%. Para a validação deste modelo, 1000 repetições foram utilizadas na construção de submodelos *jack-knife*. O número de componentes foi escolhido como antes.

A Figura 73 apresenta os resultados do modelo, onde a primeira componente mostra que todos os grupos experimentais são significativamente diferentes do grupo em 24 h, que possui o menor valor de ponderação. Isto significa que as respostas dos genes, na primeira componente, apresentadas na Figura 73B são mais importantes para os demais grupos. No entanto, observando os pesos citados na Figura 73B e os pesos dos submodelos *jack-knife* na mesma componente na Figura 73D, percebe-se uma grande diferença entre as respostas, indicando que a retirada das amostras (uma em cada grupo) provoca grande mudança no subespaço determinado pelo ML-INDSCAL para os submodelos. Provavelmente, isso se deve ao fato de que esta componente descreve vários grupos e evidencia muitas BVRs ao mesmo tempo. Adicionalmente, as diferenças, na primeira componente, entre as ponderações dos

grupos nos submodelos e aquelas do modelo global também são claras na Figura 73A (as ponderações do modelo global estão abaixo dos valores dos submodelos na maioria dos grupos), o que corrobora as afirmações acima. Todo o exposto resulta numa grande dificuldade de interpretação do modelo nesta componente.

A Figura 73E mostra que os grupos em 1 e 2 h, na segunda componente, possuem ponderações significativamente maiores que aquelas dos demais grupos, sendo importantes para esta discriminação os genes: AFAP, dentre aqueles com regulação para cima na expressão da Vpr; e o CDK10 (*cyclin-dependent kinase* (CDC2-*like*) *10*) e o CUBN (*cubilin – intrinsic factor-cobalamin receptor*), que estão entre aqueles com regulação para baixo. Outros genes também são observados nas Figuras 73B e 73F, porém estes ou não alcançam significância dentre os submodelos *jack-knife* ou não mostram importância na segunda componente do modelo ML-INDSCAL.

O gene CDK10 codifica uma proteína da subfamília CDK da família de proteínas quinases Serina/Treonina, essenciais para progressão do ciclo celular. Particularmente a proteína CDK10 desempenha um papel importante na proliferação celular, com função limitada no ciclo celular à fase G2/M [143,144], exatamente a fase em que a Vpr interrompe o processo, com o intuito de impor a replicação viral [119]. O gene CUBN está envolvido em vários processos biológicos, tais como, processo metabólico do colesterol, atuação como cobalamina e transportador de lipoproteínas, atuação como receptor e mediador na endocitose (processo pelo qual as células vivas ativamente absorvem materiais através da membrana celular) e na homeostase de tecidos [137,145]. Esta última função está provavelmente envolvida na infecção pelo HIV, pois o processo homeostático inclui o controle da proliferação celular e o reparo de tecidos, o que reflete a relação positiva do gene CUBN com CDK10. O uso de outras funções do CUBN, no contexto da infecção por HIV, não é descartada, porém, até onde se sabe não há evidências na literatura do envolvimento

deste gene na infecção citada. Conforme já comentado, o gene AFAP se relaciona com as modificações e remodelagens do citoesqueleto de actina durante a progressão celular.

Andersen & Planelles [146] comentam que quatro funções são associadas à Vpr, sendo elas: a transativação da cadeia repetida de DNA do HIV (LTR) e certos promotores heterólogos; o importe nuclear de complexos de pré-integração; a indução da interrupção do ciclo celular na fase G2/M; e a indução de apoptose (morte celular programada) em células infectadas. Destas funções, os modelos ML-INDSCAL evidenciaram elementos para a indução da interrupção do ciclo celular e para a transativação da LTR. Naturalmente, a exploração de distintas mutações como grupos de controle possibilita o foco em diferentes funcionalidades da proteína Vpr, o que explica os resultados diferentes dos modelos.

Tabela 10: Descrição dos genes estudados neste trabalho [137,143].

Gene	Processo biológico ou função molecular
TARBP1	HIV-1, o agente causador da AIDS, contém um RNA genômico que produz um DNA cromossomicamente integrado durante o ciclo replicativo. A ativação da expressão gênica do HIV-1 pelo transativador Tat é dependente do elemento regulatório de RNA (TAR) localizado abaixo do sítio de iniciação transcricional. Este elemento forma uma estrutura estável em laço e pode estar ligado ou à proteína codificada por este gene ou pela RNA polimerase II. Esta proteína pode atuar liberando a RNA polimerase II do TAR durante a elongação transcricional.
HSPC023	Pode ter um papel na hipercalcemia da malignidade (por similaridade).
MTMR1	Lipídio fosfatase que atua sobre o fosfatidilinositol 3-fosfatase e fosfatidilinositol (3,5)-bifosfato. Envolvido na rota relacionada ao metabolismo celular (metabolismo da frutose e da manose).
TRIM65	A proteína codificada contém um domínio com "cabeça" em garfo que interage com o DNA. Esta proteína liga-se às porções de purinas das cadeias repetidas de DNA do HIV (LTR), e a porções similares das purinas no promotor de interleucina 2 (IL2). Pode estar envolvida na regulação viral e de elementos promotores celulares.
GMFB	Causa a diferenciação de células cerebrais, estimula a regeneração neural, e inibe a proliferação de células tumorais. É ligante de actina e ativador enzimático. Atua como fator de crescimento, inibidor de proteína quinase e transdutor de sinal.
AFAP	Realiza ligações cruzadas com filamentos de actina dentro de estruturas em rede e em feixes. Pode modular mudanças na integridade dos filamentos de actina e induzir a formação de lamellipodia. Pode funcionar como uma molécula adaptadora que liga outras proteínas, tais como, SRC e PKC ao citoesqueleto de actina. Desempenha um papel no desenvolvimento do adenocarcinoma da próstata pela regulação da adesão célula-matriz extracelular e migração de células cancerosas.
CDK10	Catálise da reação: ATP + a proteína = ADP + a fosfoproteína. Atua na regulação negativa da proliferação celular.
CUBN	Co-transportador que desempenha um papel no metabolismo do ferro, vitaminas e lipoproteínas, facilitando a absorção. Liga-se a ALB, MB, Kappa e cadeias lambda, TF, hemoglobina, GC, SCGB1A1, APOA1, lipoproteínas de alta densidade e ao complexo GIF-cobalamina, requerendo cálcio. Serve como transportador em epitélios absortivos como, intestino, tubos renais e saco vitelino embrionário. A interação com LRP2 media seu tráfego através das vesículas e facilita a absorção de ligantes específicos (GC, hemoglobina, ALB, TF e SCGB1A1). A interação com AMN controla seu tráfego para a membrana plasmática e facilita a endocitose de ligantes. Pode desempenhar um papel no desenvolvimento da peri-implantação do embrião através da internalização de APOA1 e colesterol. Liga-se ao LGALS3 na interface maternal-fetal.

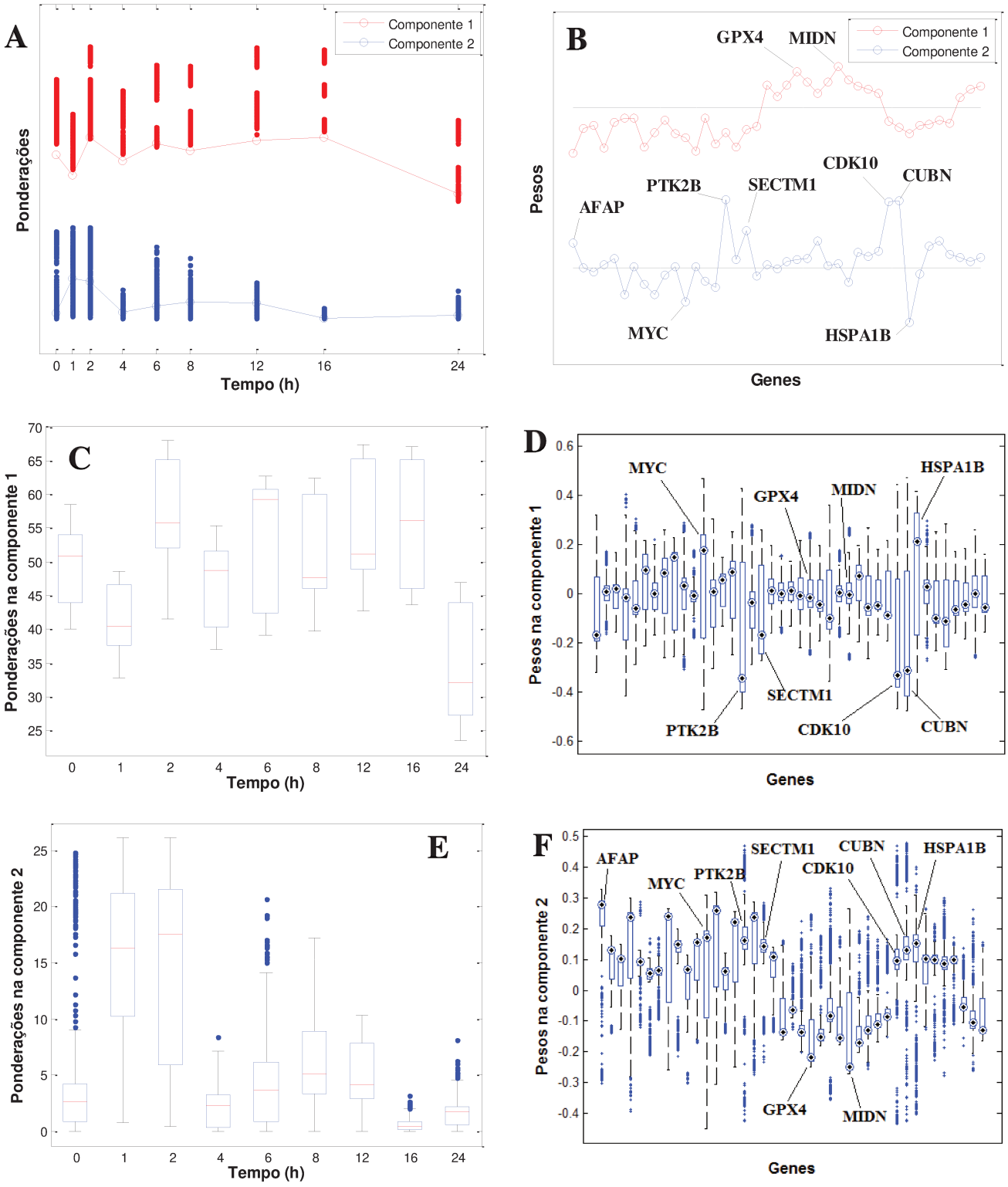

Figura 73: Modelo ML-INDSCAL com dois componentes para os 42 genes expressos especificamente para a Vpr na forma nativa comparando com a mutação R80A. A: Ponderações para os grupos experimentais e submodelos *jack-knife* (pontos coloridos); B: Pesos para as variáveis (genes), a linha cinza indica o zero; Gráficos de

caixa obtidos através das estatísticas dos submodelos *jack-knife*. C: Ponderações na componente 1; D: Pesos na componente 1; E: Ponderações na componente 2; F: Pesos na componente 2. Em A e B os valores estão deslocados verticalmente para uma melhor visualização e suas escalas são similares aos respectivos gráficos em caixas.

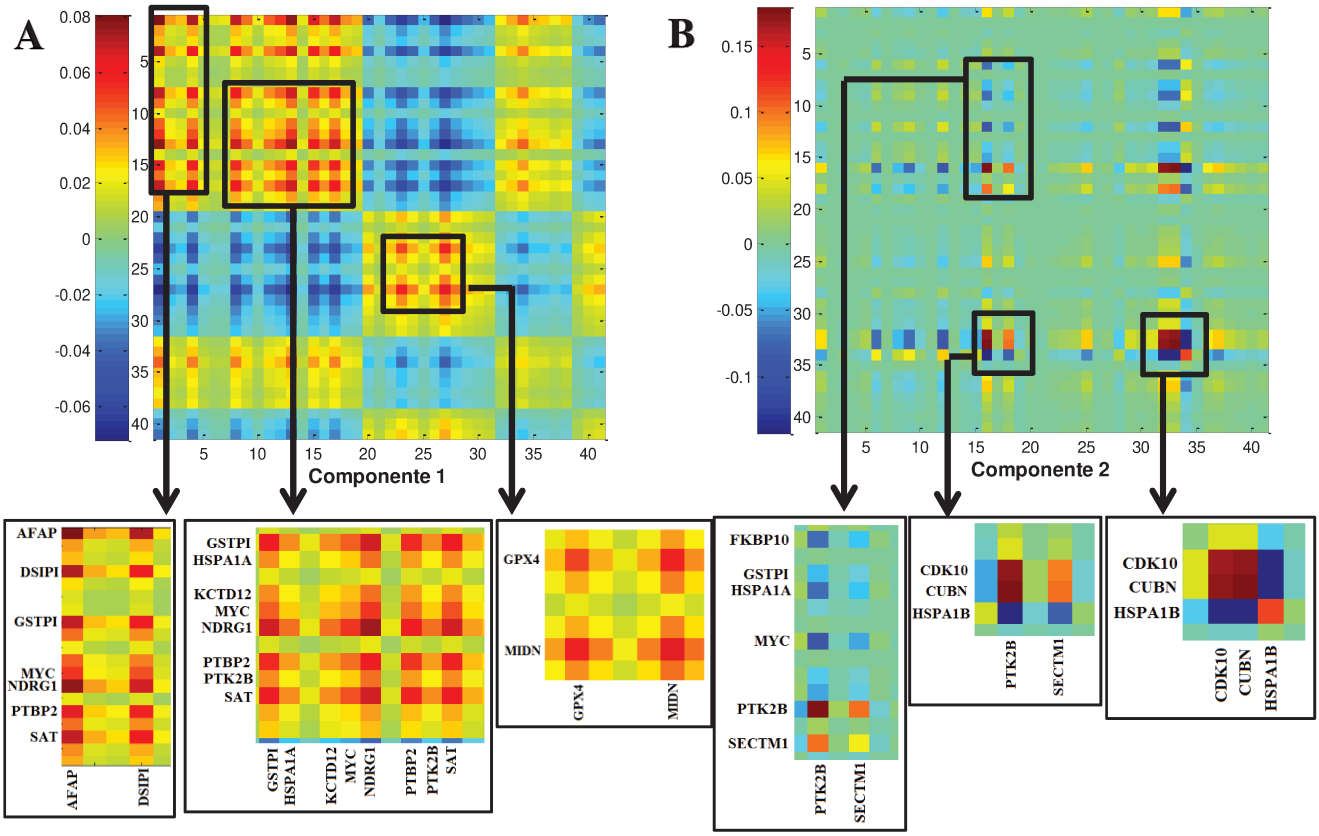

Figura 74: Gráficos de calor obtidos a partir dos pesos do modelo ML-INDSCAL na Figura 73. A: Componente 1; B: Componente 2. Em cada gráfico, regiões ampliadas com os genes mais importantes são mostradas.

4.6.2. Conjunto de dados real: comparação com modelos PCA

Modelos PCA foram construídos para as matrizes de dados, centradas na média, estudadas por ML-INDSCAL e são apresentados nas Figuras 75 e 76. O modelo da Figura 75 diz respeito aos dados que compararam as amostras na forma nativa com aquelas das formas com mutação F72A/R73A, enquanto o modelo da Figura 76 corresponde à comparação com a mutação R80A como grupo de controle.

Em ambos os modelos pode ser visto que a principal variância, capturada na primeira componente principal, fornece a diferenciação entre as linhas nativas e as linhas com mutação, basicamente devido às respostas dos genes com regulação para cima e para baixo [119], como pode ser notado nos pesos em PC1 nas Figuras 75B e 76B. Ainda em PC1, a razão para as linhas com mutação possuírem escores significativamente menores está relacionado com o fato de que os conjuntos de dados são compostos por genes caracterizados como eventos específicos das formas nativas e assim os grupos de controle tendem a possuir menor resposta de expressão genética para as variáveis com regulação para cima e maior resposta para aquelas com regulação para baixo. Por outro lado, os modelos ML-INDSCAL considerando as diferenças entre as linhas celulares foca na variação intra-individual e assim tende a remover a variação citada (variação estática). A PCA geralmente confunde estas variações o que pode dificultar as interpretações [112].

Observando os gráficos de escores em PC2 e PC3 no modelo da Figura 75 que envolve as linhas celulares com mutação F72A/R73A como grupo de controle e comparando os escores médios em cada grupo experimental (grupos no controle e na linha nativa), cujas variabilidades são consideradas pelos escores das amostras individuais, pode ser notada a diferença significativa dos grupos de amostras expressando a Vpr nativa em 1, 4 e 16 h em PC2 e PC3. As Figuras 75C e 75D mostram que em PC2 e PC3 as respostas incluem principalmente os mesmos genes

apontados pelo modelo ML-INDSCAL, mas com distintas BVRs. As componentes do modelo ML-INDSCAL descrevem as BVRs nos grupos em 4 e 16 h somente na primeira componente, separadamente das BVRs no grupo experimental em 1 h, descritas na segunda componente. Por outro lado, o modelo PCA descreve parte das relações no grupo em 4 h em PC3 e inclui as respostas para o grupo em 1 h em PC2. Portanto, o modelo ML-INDSCAL apresenta resultados mais parcimoniosos e assim considerados mais confiáveis. Isto pode ser explicado pelos diferentes focos dos métodos, onde enquanto a PCA foca nos níveis das variáveis, o ML-INDSCAL foca diretamente nas variâncias e covariâncias e assim descrevendo as BVRs mais fielmente.

O modelo PCA apresentado na Figura 76 não atinge diferenças significativas nem entre os escores das linhas celulares nativas e aqueles para as linhas celulares com mutação R80A nem entre os grupos experimentais, em PC2 e PC3, apesar de indicar a importância de genes também indicados pelo modelo ML-INDSCAL. Neste exemplo, provavelmente a presença da variação não dinâmica dificulta a evidencia das diferenças. Pode ser observado que as variações capturadas em PC2 e PC3 mostram respostas para os grupos em 0, 1, 2, 3 e 4 h, no conjunto de dados do grupo controle (linhas com mutação R80A), que no modelo ML-INDSCAL são consideradas como variação estática. De fato, o papel do grupo controle é exatamente fornecer a variação que deve ser desconsiderada como induzida pelo tratamento.

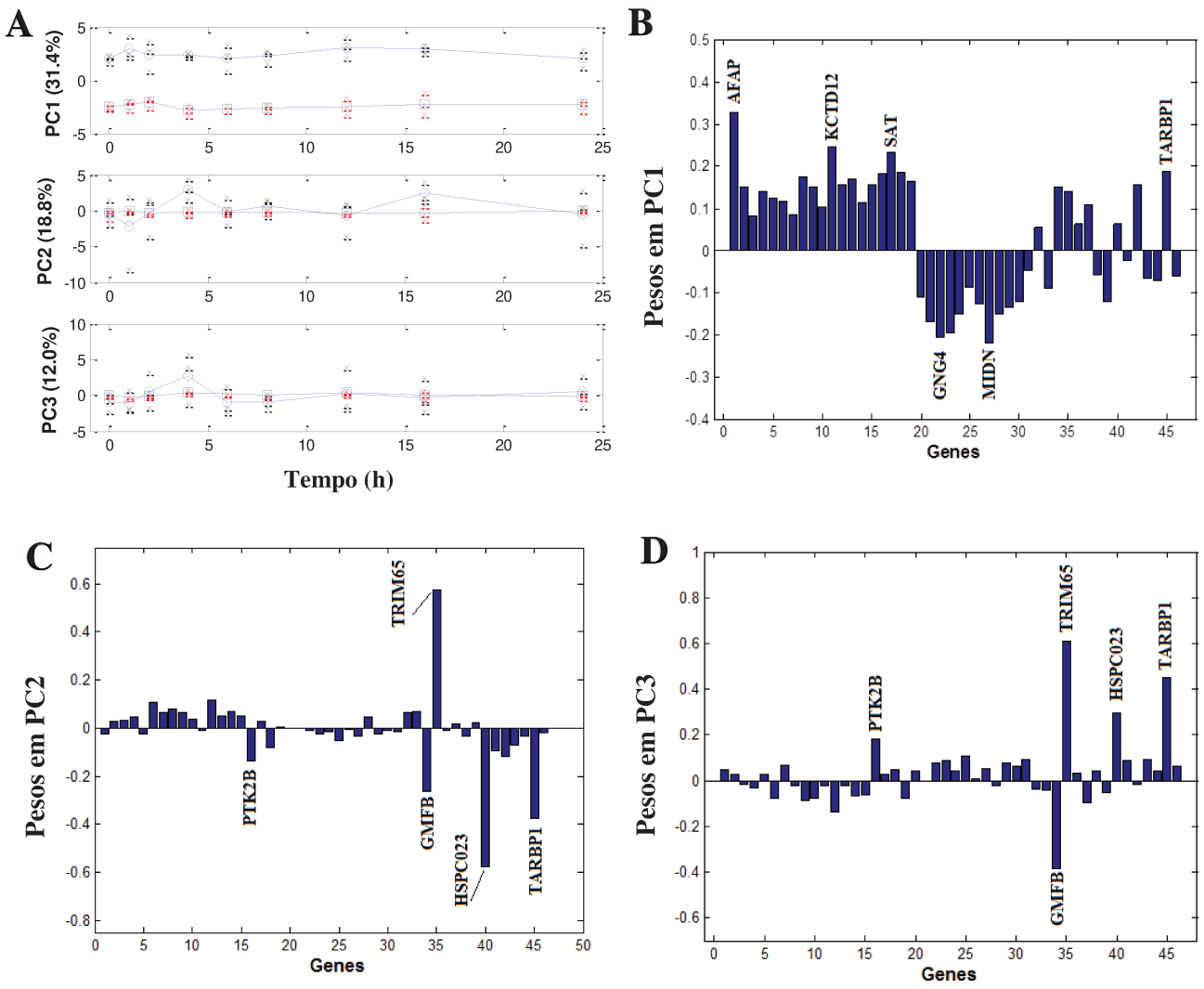

Figura 75: Modelo PCA (três primeiras componentes principais) para o conjunto de dados com os 47 genes das linhas celulares nativas e com mutação F72A/R73A. A: Gráfico de escores: círculos azuis representam as médias dos escores em cada grupo da linha nativa e quadrados azuis representam as médias dos escores em cada grupo da linha com mutação. As cruzes pretas representam os escores individuais da linha nativa, enquanto as cruzes vermelhas representam os escores individuais da linha com mutação; B: Gráfico de pesos para a primeira componente principal; C: Gráfico de pesos para a segunda componente principal; D: Gráfico de pesos para a terceira componente principal.

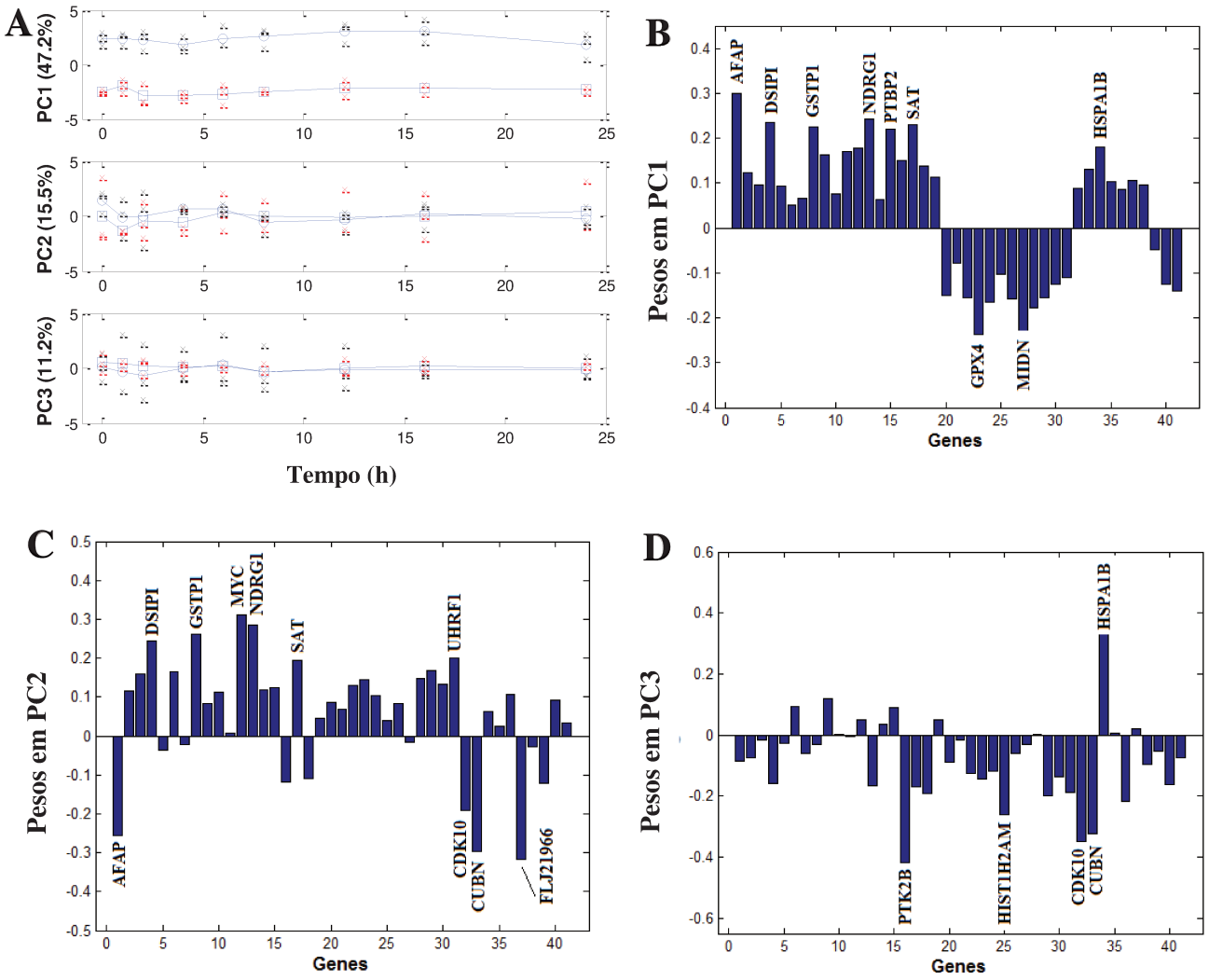

Figura 76: Modelo PCA (três primeiras componentes principais) para o conjunto de dados com os 47 genes das linhas celulares nativas e com mutação F72A/R73A. A: Gráfico de escores: círculos azuis representam as médias dos escores em cada grupo da linha nativa e quadrados azuis representam as médias dos escores em cada grupo da linha com mutação. As cruzes pretas representam os escores individuais da linha nativa, enquanto as cruzes vermelhas representam os escores individuais da linha com mutação; B: Gráfico de pesos para a primeira componente principal; C: Gráfico de pesos para a segunda componente principal; D: Gráfico de pesos para a terceira componente principal.

4.7. Conclusões

Neste trabalho, apresentou-se o método ML-INDSCAL baseado na ideia de separação da variação, para isolar a variação intra-individual, através da subtração de respostas específicas (respostas no primeiro ponto do tempo ou de um grupo de controle), seguida da modelagem das matrizes de covariância, computadas de cada grupo experimental. As subtrações são realizadas de maneira diferente dependendo do tipo de planejamento experimental no qual a aquisição dos dados se fundamenta (planejamento cruzado ou paralelo). A modelagem usa PARAFAC para determinar o subespaço onde cada grupo experimental tem uma ponderação positiva. Grupos com valores mais altos de ponderação são interpretados como aqueles onde as relações mostradas no subespaço (pesos) são mais importantes. Este é um novo modo de verificar diferenças entre grupos experimentais no curso do tempo. Tais diferenças são ditadas pelas mudanças simultâneas nas variâncias e covariâncias. Ao final, os modelos são validados por uma versão do procedimento *jack-knife*, um método de re-amostragem que aqui não é baseado em suposições sobre as distribuições das estatísticas calculadas.

Explorando um simples exemplo simulado, mostrou-se o desempenho do método ML-INDSCAL para tratar um conjunto de dados onde as covariâncias mudam drasticamente (com inversão de sinal) entre dois grupos experimentais, sendo a diferenciação exatamente baseada neste fato. As comparações com os métodos PCA, INDSCAL e PRC evidenciaram a melhor solução dada pelo ML-INDSCAL, uma vez que este foi capaz de diferenciar os grupos experimentais simulados entre si e entre um grupo de controle simulado.

Considerando um conjunto de dados real de um estudo de perfis de expressão genética celular global, onde linhas celulares expressando a proteína Vpr na forma nativa e duas linhas onde a Vpr tem mutações (F72A/R73A e R80A) foram

investigadas, o método ML-INDSCAL foi capaz de evidenciar funções distintas da proteína Vpr através das BVRs dentro dos grupos experimentais. Comparando os modelos ML-INDSCAL com modelos PCA, a abordagem aqui apresentada mostrou modelos mais parcimoniosos e bem sucedidos na descrição das relações entre os genes.

Acredita-se que o método ML-INDSCAL, dada as suas características citadas, pode ser um método complementar no campo da análise de dados temporais, lançando nova luz sobre o campo de seleções de características. A ML-INDSCAL toolbox composta por rotinas escritas em código Matlab encontra-se disponível para a realização de todas as tarefas citadas neste trabalho e pode ser adquirida gratuitamente nos sítios de internet: http://lqta.iqm.unicamp.br ou http://www.bdagroup.nl.

Considerações finais do trabalho

Neste trabalho, duas novas metodologias de análise multivariada foram apresentadas. A primeira delas consistiu no desenvolvimento de uma ferramenta auxiliar, denominada *bucketing* otimizado, para resolver os problemas de desalinhamentos nos espectros de ressonância magnética nuclear (RMN) de próton. Esta abordagem foi apresentada com estudos de caso, onde as suas vantagens e desvantagens foram discutidas. Conforme mostrado a solução fornecida pela ferramenta é satisfatória, fornecendo valiosa ajuda para a construção de modelos de calibração (estudo de caso de vinhos), de classificação (estudo de caso de tumores cerebrais) e de análise exploratória (estudo de caso de amostras de biodiesel-diesel). A análise de componentes principais em intervalos (iPCA) foi aplicada para o controle de qualidade de amostras de biodiesel-diesel, através da exploração dos espectros de RMN de (^{1}H) próton e de (^{13}C) carbono.

Para os espectros de carbono, um grande problema residiu no fato dos mesmos se apresentarem bastantes ruidosos, necessitando de uma abordagem para diminuição dos ruídos. Assim, a análise de componentes principais em múltiplas escalas (MSPCA) foi empregada, utilizando as vantagens da transformada *wavelet* e da PCA. O método mostrou-se eficiente, porém como envolve muitos fatores (escolha do número de níveis de decomposição na transformada *wavelet*, escolha das funções de base, escolha dos números de componentes principais nas etapas de PCA, entre outros) não pode ser considerado um método prático, caso necessite ser utilizado em análises de rotina, por exemplo. Este fato é um dos quais torna o tratamento dos espectros de carbono menos vantajoso frente ao tratamento dos espectros de próton.

Os modelos iPCA utilizando os espectros de RMN ^{1}H das duas classes, metropolitana e não metropolitana, analisadas em conjunto e separadamente, mostraram os melhores desempenhos. Assim, sugere-se que na análise por esta

metodologia, os dois tipos de modelo sejam usados de forma complementar. Ainda sobre estes modelos, vale ressaltar o ponto fraco observado em relação às amostras ligeiramente fora da especificação para o teor de biodiesel. Para este parâmetro, um ponto positivo para os modelos discutidos foi a observação de que amostras com teores bem diferentes foram prontamente identificadas.

Conforme foi discutido, o problema citado talvez possa ser contornado utilizando um modelo com um número maior de amostras conformes para o teor de biodiesel. Por outro lado, neste contexto, há a possibilidade do perfil espectral naturalmente não poder expressar as diferenças das amostras não conformes. Uma solução para esta questão poderia ser a consideração de um aumento na faixa ao redor do valor do parâmetro exigido na legislação, permitindo que haja uma maior diferença entre o padrão espectral de uma amostra reprovada e uma não reprovada.

Felizmente, uma amostra ligeiramente fora da especificação para o teor de biodiesel não necessariamente representa um biocombustível com uma qualidade ruim, apesar de não atender a legislação. Ao longo dos anos o teor de biodiesel tem sido inclusive variável, sendo que o governo Brasileiro vem estendendo gradualmente a porcentagem em volume de biodiesel na mistura comercializada nos postos.

Na prática, certa atenção deve ser dada às amostras detectadas pelo modelo e que não são reprovadas pelos parâmetros físico-químicos. Considerando-se que a aquisição dos espectros é adequada, as diferenças detectadas nos perfis espectrais destas amostras podem sugerir a ocorrência de adulterações que não são identificadas pelos métodos de rotina.

Na segunda metodologia proposta neste trabalho, o novo método denominado escalamento de diferenças individuais multinível (ML-INDSCAL) apresenta-se como uma abordagem promissora, uma vez que considera efeitos nos dados que não são contemplados pelos métodos padrões como PCA, PLS-DA, entre outros.

As mudanças nos sistemas biológicos são bastante complexas e nem sempre ocorrem de modo semelhante para todos os indivíduos estudados. Desse modo, as mudanças nas estruturas de correlação (ou covariância) são muito comuns, podendo representar o efeito de algum tratamento. Além disso, nem todas as mudanças num sistema biológico são decorrentes do tratamento, ou seja, um tratamento atinge apenas uma parte das rotas biológicas. As demais continuam suas modificações independentes do tratamento, representando a variação biológica normal do indivíduo.

O método aqui proposto combina a exploração isolada da variação induzida pelo tratamento, denominada variação intra-individual, com a busca por mudanças nas estruturas de covariância, sendo, portanto, um método novo para os dados em ciências ômicas.

Demonstrou-se através de um exemplo simulado e um exemplo real a força do método e espera-se a partir disto que o mesmo ganhe espaço dentro do campo de seleção de características, como uma abordagem capaz de revelar biomarcadores não antes explorados em certos sistemas biológicos.

Referências Bibliográficas

[1] Chemometric Consultancy [Online]. "http://www.chemometry.com/Index/Chemometrics.html". Acesso em 05 de Junho de 2013, às 14h.

[2] R Kiralj & MMC Ferreira, "The past, present, and future of chemometrics worldwide: some etymological, linguistic, and bibliometric investigations," *Journal of Chemometrics*, vol. 20, pp. 247–272, 2006.

[3] C-J Xu, HCJ Hoefsloot, AK Smilde, "To aggregate or not to aggregate high-dimensional classifiers," *BMC Bioinformatics*, vol. 12, pp. 153-160, 2011.

[4] S Wold, K Esbensen, P Geladi, "Principal Component Analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, pp. 37-52, 1987.

[5] K Esbensen & P Geladi, "Principal component analysis: concept, geometrical interpretation, mathematical background, algorithms, history, practice.," in *Comprehensive Chemometrics*, Stephen D Brown, Romà Tauler, and Beata Walczak, Eds. Amsterdam, The Netherlands: Elsevier B. V., 2009, ch. 2.13, pp. 211-226.

[6] SK Booksh & RB Kowalski, "Theory of analytical chemistry," *Analytical Chemistry*, vol. 66, pp. 782-791, Agosto 1994.

[7] R Bro, *Multi-Way Analysis in the Food Industry. Models, Algorithm and Applications*. Amsterdam: University of Amsterdam, 1998, Tese de doutorado.

[8] S Liu & G Trenkler, "Hadamard, Khatri-Rao, Kronecker and other matrix products," *International Journal of information and systems sciences*, vol. 4, pp. 160-177, 2008.

[9] G Tomasi & R Bro, "Multilinear Models: Iterative Methods," in *Comprehensive Chemometrics*, Stephen D Brown, Romà Tauler, and Beata Walczak, Eds. Amsterdam, The Netherlands: Elsevier B. V., 2009, ch. 2.22, pp. 411-451.

[10] R Bro, "Parafac. Tutorial and applications," *Chemomemcs and Intelligent Laboratory Systems*, vol. 38, pp. 149-171, 1997.

[11] R Bro & HAL Kiers, "A new efficient method for determining the number of components in PARAFAC models," *Journal of Chemometrics*, vol. 17, pp. 274-286, 2003.

[12] J Riu & R Bro, "Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models," *Chemometrics and Intelligent Laboratory Systems*, vol. 65, pp. 35-49, 2003.

[13] MM Sena, MG Trevisan, RJ Poppi, "Parafac: uma ferramenta quimiométrica para tratamento de dados multidimensionais. Aplicação na determinação direta

de fármacos em plasma humano por espectrofluorimetria," *Química Nova*, vol. 28, pp. 910-920, 2005.

[14] CA Andersson & R Bro, "The N-way Toolbox for MATLAB," *Chemometrics and Intelligent Laboratory Systems*, vol. 52, pp. 1-4, 2000.

[15] H Abdollahi & SM Sajjadi, "On rotational ambiguity in parallel factor analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 103, pp. 144-151, 2010.

[16] MR Monteiro, ARP Ambrozin, MS Santos, EF Boffo, ER Pereira Filho, LM Lião, AG Ferreira, "Evaluation of biodiesel-diesel blends quality using 1H NMR and chemometrics," *Talanta*, vol. 78, pp. 660-664, 2009.

[17] SAA Sousa, A Magalhães, MMC Ferreira, "Optimized bucketing for NMR spectra: Three cases studies," *Chemometrics and Intelligent Laboratory Systems*, vol. 122, pp. 93-102, 2013.

[18] JS Ribeiro, F Augusto, TJG Salva, MMC Ferreira, "Prediction models for Arabica coffee beverage quality based on aroma analyses and chemometrics," *Talanta*, vol. 101, pp. 253-260, 2012.

[19] EB Melo, JPA Martins, TCM Jorge, MC Friozi, MMC Ferreira, "Mutivariate QSAR study on the antimutagenic activity of flavonoids against 3-NFA on Salmonella typhimurium TA98," *European Journal of Medicinal Chemistry*, vol. 45, pp. 4562-4569, 2010.

[20] GH Golub & CF van Loan, *Matrix computation*. Baltimore: John Hopkins University Press, 1996.

[21] JPA Martins, RF Teófilo, MMC Ferreira, "Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets," *Journal of Chemometrics*, vol. 24, pp. 320-332, 2010.

[22] R Andersson, A comparison of nine PLS1 algorithms. *Journal of Chemometrics*, vol. 23, pp. 518-529, 2009.

[23] H Martens & T Naes, *Multivariate Calibration*. New York: Wiley, 1989.

[24] P Geladi & BR Kowalski, "Partial least-squares regression - A tutorial," *Analytica Chimica Acta*, vol. 185, pp. 1-17, 1986.

[25] M Maeder & YM Neuhold, *Data Handling in Science and Technology: Practical Data Analysis in Chemistry*, S Rutan and B Walczak, Eds. Amsterdam: Elsevier B. V., 2007, vol. 26.

[26] R Kiralj & MMC Ferreira, "Basic validation procedures for regression models in QSAR and QSPR studies: theory and application," *Journal of Brazilian Chemical Society*, vol. 20, pp. 770-787, 2009.

[27] JS Ribeiro, F Augusto, TJG Salva, RA Thomaziello, MMC Ferreira,

"Prediction of sensory properties of Brazilian Arabica roasted coffees by headspace solid phase microextraction-gas chromatography and partial least squares," *Analytica Chimica Acta*, vol. 634, pp. 172-179., 2009.

[28] EJJ van Velzen, JA Westerhuis, JPM van Duynhoven, FA van Dorsten, HCJ Hoefsloot, DM Jacobs, S Smit, R Draijer, CI Kroner, AK Smilde, "Multilevel data analysis of a crossover designed human nutritional intervention study," *Journal of Proteome Research*, vol. 7, pp. 4483-4491, 2008.

[29] EJJ van Velzen, JA Westerhuis, JPM van Duynhoven, FA van Dorsten, CH Grün, DM Jacobs, GSMJE Duchateau, DJ Vis, AK Smilde, "Phenotyping tea consumers by nutrikinetic analysis of polyphenolic end-metabolites," *Journal of Proteome Research*, vol. 8, pp. 3317-3330, 2009.

[30] E Szymanska, E Saccenti, AK Smilde, JA Westerhuis, "Double-check: validation of diagnostic statistics for PLS-DA models in metabolomic studies," *Metabolomics*, vol. 8, no. Suppl 1, pp. 3-16, 2012.

[31] JA Westerhuis, EJJ van Velzen, HCJ Hoefsloot, AK Smilde, "Discriminant Q2 (DQ2) for improved discrimination in PLSDA models," *Metabolomics*, vol. 4, pp. 293-296, 2008.

[32] JA Westerhuis, HCJ Hoefsloot, S Smit, DJ Vis, AK Smilde, EJJ van Velzen, JPM van Duynhoven, FA van Dorsten, "Assessment of PLSDA cross validation," *Metabolomics*, vol. 4, pp. 81–89, 2008.

[33] AV Faria, FC Macedo Jr, AJ Marsaioli, MMC Ferreita, F Cendes, "Classification of brain tumor extracts by high resolution 1H MRS using partial least squares discriminant analysis," *Brazilian Journal of Medical and Biological Research*, vol. 44, pp. 149-164, 2011.

[34] NF Pérez, J Ferré, R Boqué, "Calculation of the reliability of classification in discriminant partial least-squares binary classification," *Chemometrics and Intelligent Laboratory Systems*, vol. 95, pp. 122-128, 2009.

[35] Eigenvector Research Incorporated. How is the prediction probability and threshold calculated for PLSDA? [Online]. "http://www.eigenvector.com/faq/?id=38". Acesso em 10 de Junho de 2013, às 16h.

[36] EM Lenz, J Bright, ID Wilson, SR Morgan, AFP Nash, "A ^{1}H NMR-based metabonomic study of urine and plasma samples obtained from healthy human subjects," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 33, pp. 1103-1115, 2003.

[37] S Agnolet, JW Jaroszewski, R Verpoorte, D Staerk, "^{1}H NMR-based metabolomics combined with HPLC-PDA-MSSPE-NMR for investigation of standardized *Ginkgo biloba* preparations," *Metabolomics*, vol. 6, pp. 292–302,

2010.

[38] EF Boffo, LA Tavares, MMC Ferreira, AG Ferreira, "Classification of Brazilian vinegars according to their 1H NMR spectra by pattern recognition analysis," *LWT- Food Science and Technology*, vol. 42, pp. 1455-1460, 2009.

[39] MR Monteiro, ARP Ambrozin, LM Lião, EF Boffo, LA Tavares, MMC Ferreira, AG Ferreira, "Study of brazilian gasoline quality using hydrogen nuclear magnetic resonance (1H NMR) spectroscopy and chemometrics," *Energy & Fuels*, vol. 23, pp. 272-279, 2009.

[40] C Daolio, FL Beltrame, AG Ferreira, QB Cass, DA Cortez, MMC Ferreira, "Classification of commercial catuaba samples by NMR, HPLC and chemometrics," *Phytochemical Analysis*, vol. 19, pp. 218-228, 2008.

[41] RH Jellema, "Variable shift and alignment," in *Comprehensive Chemometrics*, S D Brown, R Tauler, and B Walczak, Eds. Amsterdam, The Netherlands: Elsevier, 2009, ch. 2.06, pp. 85-108.

[42] H Winning, *Quantitative multivariate NMR spectroscopy in Food Science and Nutrition*. Frederiksberg, Denmark: University of Copenhagen, 2009, Tese de doutorado.

[43] N Trbovic, F Dancea, T Langer, U Günther, "Using wavelet de-noised spectra in NMR screening," *Journal of Magnetic Resonance*, vol. 173, pp. 280-287, 2005.

[44] J Forshed, RJ Torgrip, KM Aberg, B Karlberg, J Lindberg, SP Jacobsson, "A comparison of methods for alignment of NMR peaks in the context of cluster analysis," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 38, pp. 824-832., 2005.

[45] V Pravdova, B Walczak, DL Massart, "A comparison of two algorithms for warping of analytical signals," *Analytical Chimica Acta*, vol. 456, pp. 77-92, 2002.

[46] T Skov, F van der Berg, G Tomasi, R Bro, "Automated alignment of chromatographic data," *Journal of Chemometrics*, vol. 20, pp. 484-497, 2006.

[47] G Tomasi , F van der Berg, C Andersson, "Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data," *Journal of Chemometrics*, vol. 18, pp. 231-241, 2004.

[48] NPV Nielsen, JM Carstensen, J Smedsgaard, "Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimized warping," *Journal of Chromatography*, vol. 805, pp. 17-35, 1998.

[49] FH Larsen, F van der Berg, SB Engelsen, "An exploratory chemometric study of ^{1}H NMR spectra of table wines," *Journal of Chemometrics*, vol. 20, pp. 198-

208, 2006.

[50] F Savorani, G Tomasi, SB Engelsen, "icoshift: A versatile tool for the rapid alignment of 1D NMR spectra," *Journal of Magnetic Resonance*, vol. 202, pp. 190-202, 2010.

[51] B Lefebvre, R Sasaki, S Golotvin, AW Nicholls, Intelligent bucketing for metabonomics – Part 1. [Online]. "http://www.acdlabs.com/download/publ/2004/enc04/intelbucket.pdf/". Acesso em 12 de Janeiro de 2013, às 12h.

[52] B Lefebvre, R Sasaki, S Golotvin, AW Nicholls, Intelligent bucketing for metabonomics – Part 2. [Online]. "http://www.acdlabs.co.uk/download/publ/2004/intelbucket2.pdf/". Acesso em 12 de Janeiro de 2013, às 12h.

[53] AW Nicholls, RJ MortiShire-Smith, JK Nichols, "NMR spectroscopic-based metabonomic studies of urinary metabolite variation in acclimatizing germ-free rats," *Chemical Research in Toxicology*, vol. 16, pp. 1395-1404, 2003.

[54] L Jang-Eun, H Geum-Sook, VDB Frans, L Cherl-Ho, H Young-Shick, "Evidence of vintage effects on grape wines using 1H NMR-based metabolomic study," *Analytica Chimica Acta*, vol. 648, pp. 71–76, 2009.

[55] RA Davis, AJ Charlton, J Godward, SA Jones, M Harrison, JC Wilson, "Adaptive binning: An improved binning method for metabolomics data using the undecimated wavelet transform," *Chemometrics and Intelligent Laboratory Systems*, vol. 85, pp. 144-154, 2007.

[56] PE Anderson, NV Reo, NJ DelRaso, TE Doom, ML Raymer, "Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics," *Metabolomics*, vol. 4, pp. 261-272, 2008.

[57] PE Anderson, DA Mahle, TE Doom, NV Reo, NJ DelRaso, ML Raymer, "Dynamic adaptive binning: an improved quantification technique for NMR spectroscopic data," *Metabolomics*, vol. 7, pp. 179-190, 2011.

[58] KM Åberg, E Alm, RJO Torgrip, "The correspondence problem for metabonomics datasets," *Analytical and Bioanalytical Chemistry*, vol. 394, pp. 151-162, 2009.

[59] E Alm, RJO Torgrip, KM Åberg, I Schuppe-Koistinen, J Lindberg, "A solution to the 1D NMR alignment problem using an extended generalized fuzzy Hough transform and mode support," *Analytical and Bioanalytical Chemistry*, vol. 395, pp. 213-223, 2009.

[60] E Alm, T Slagbrand, KM Åberg, E Wahlström, I Gustafsson, J Lindberg, "Automated annotation and quantification of metabolites in ^{1}H NMR data of biological origin," *Analytical and Bioanalytical Chemistry*, vol. 403, pp. 443-

455, 2012.

[61] M Insausti, C Romano, MF Pistonesi, BSF Band, "Simultaneous determination of quality parameters in biodiesel/diesel blends using synchronous fluorescence and multivariate analysis," *Microchemical Journal*, vol. 108, pp. 32-37, 2013.

[62] G Knothe, "Biodiesel and renewable diesel: A comparison," *Progress in Energy and Combustion Science*, vol. 36, pp. 364-373, 2010.

[63] G Knothe, "Dependence of biodiesel fuel properties on the structure of fatty acid alkyl esters," *Fuel Processing Technology*, vol. 86, pp. 1059-1070, 2005.

[64] G Knothe, AC Matheaus, TW Ryan III, "Cetane numbers of branched and straight-chain fatty esters determined in an ignition quality tester," *Fuel*, vol. 82, pp. 971-975, 2003.

[65] AK Agarwal, "Biofuels (alcohols and biodiesel) applications as fuel for internal combustion engines," *Progress in Energy and Combustion Science*, vol. 33, pp. 233-271, 2007.

[66] A Talebian-Kiakalaieh, NAS Amin, H Mazaheri, "A review on novel processes of biodiesel production from waste cooking oil," *Applied Energy*, vol. 104, pp. 683–710, 2013.

[67] DYC Leung, X Wu, MKH Leung, "A review on biodiesel production using catalyzed transesterification," *Applied Energy*, vol. 87, pp. 1083–1095, 2010.

[68] W-B Zhang, "Review on analysis of biodiesel with infrared spectroscopy," *Renewable and Sustainable Energy Reviews*, vol. 16, pp. 6048–6058, 2012.

[69] AC Pinto, LLN Guarieiro, MJC Rezende, NM Ribeiro, EA Torres, WA Lopes, PAP Pereira, JB Andrade, "Bioidiesel: An overview," *Journal of Brazilian Chemical Society*, vol. 16, pp. 1313-1330, 2005.

[70] LC Meher, VD Sagar, SN Naik, "Technical aspects of biodiesel production by transterification - a review," *Renewable and Sustainable Energy Reviews*, vol. 10, pp. 248-268, 2006.

[71] IS Flores, MS Godinho, AE Oliveira, GB Alcântara, MR Monteiro, SMC Menezes, LM Lião, "Discrimination of biodiesel blends with ^{1}H NMR spectroscopy and principal component analysis," *Fuel*, vol. 99, pp. 40-44, 2012.

[72] AE Atabani, AS Silitonga, HC Ong, TMI Mahlia, HH Masjuki, IA Badruddin, H Fayaz, "Non-edible vegetable oils: A critical evaluation of oil extraction, fatty acid compositions, biodiesel production, characteristics, engine performance and emissions production," *Renewable and Sustainable Energy Reviews*, vol. 18, pp. 211–245, 2013.

[73] Sebrae - Serviço Brasileiro de Apoio às Micro e Pequenas Empresas.

Agroenergia - Cartilha: Biodiesel. [Online]. "http://bis.sebrae.com.br/GestorRepositorio/ARQUIVOS_CHRONUS/bds/bds.nsf/D170D324C7521915832572B200470F63/$File/NT00035116.pdf". Acesso em 10 de Junho de 2013, às 14h.

[74] RP Scherer, AL Malvesti, SBC Pergher, WF Souza, "Estudo de adsorção de compostos sulfurados utilizando um diesel comercial dopado com benzotiofeno e dibenzotiofeno," *Química Nova*, vol. 32, pp. 34-37, 2009.

[75] RT Yang, AJ Hernández-Maldonado, FH Yang, "Desulfurization of transportation fuels with zeolites under ambient conditions," *Science Reports*, vol. 301, pp. 79-81, 2003.

[76] Agência Nacional de Petróleo Gás Natural e Biocombustíveis. Resoluções ANP. [Online]. "http://nxt.anp.gov.br/NXT/gateway.dll/leg/resolucoes_anp/2012/dezembro/ranp%2046%20-%202012.xml?fn=document-frameset.htm$f=templates$3.0". Acesso em 10 de Junho de 2013, às 14h.

[77] U Rashid, F Anwar, G Knothe, "Evaluation of biodiesel obtained from cottonseed oil," *Fuel Processing Technology*, vol. 90, pp. 1157-1163, 2009.

[78] G Knothe & KR Steidley, "A comparison of used cooking oils: A very heterogenous feedstok for biodiesel," *Bioresource Technology*, vol. 100, pp. 5796-5801, 2009.

[79] G Knothe & KR Steidley, "Kinematic viscosity of biodiesel components (fatty acid alkyl esters) and related compounds at low temperatures," *Fuel*, vol. 86, pp. 2560-2567, 2007.

[80] G Knothe & KR Steidley, "Kinematic viscosity of biodiesel fuel components and related compounds. Influence of compound structure and comparison to petrodiesel fuel components," *Fuel*, vol. 84, pp. 1059-1065, 2005.

[81] G Knothe, "Some aspects of biodiesel oxidative stability," *Fuel Processing Technology*, vol. 88, pp. 669-677, 2007.

[82] G Knothe, "Determining the blend level of mixtures of biodiesel with conventional diesel fuel by fiber-optic near-infrared spectroscopy and 1H nuclear magnetic resonance spectroscopy," *Journal of American Oil Chemistry Society*, vol. 78, pp. 1025-1028, 2001.

[83] P Baptista, P Felizardo, JC Menezes, MJN Correia, "Multivariate near infrared spectroscopy models for predicting the iodine value, CFPP, kinematic viscosity at 40 ºC and density at 15 ºC of biodiesel," *Talanta*, vol. 77, pp. 144-151, 2008.

[84] MR Monteiro, ARP Ambrozin, LM Lião, AG Ferreira, "Determination of biodiesel blend levels in different diesel samples by 1H NMR," *Fuel*, vol. 88, pp. 691-696, 2009.

[85] G Knothe, MO Bagby, D Weisleder, "Evaluation of the olefinic proton signals in the 1H NMR spectra of allylic hydroxy groups in long-chain compounds," *Chemistry and Physics of Lipids*, vol. 82, pp. 33-37, 1996.

[86] M Nagy, BJ Kerr, CJ Ziemer, AJ Ragauskas, "Phosphitylation and quantitativa 31P NMR analysis of partially substituted biodiesel glycerols," *Fuel*, vol. 88, pp. 1793-1797, 2009.

[87] J Keeler. Understanding NMR spectroscopy. [Online]. "http://www-keeler.ch.cam.ac.uk/lectures". Acesso em 08 de Maio de 2009, às 14h.

[88] AMK Pedro & MMC Ferreira, "Simultaneously calibrating solids, sugars and acidity of tomato products using PLS2 and NIR spectroscopy," *Analytica Chimica Acta*, vol. 595, pp. 221-227, 2007.

[89] N Serban, "Noise reduction for enhanced component identification in multi-dimensional biomolecular NMR studies," *Computational Statistics and Data Analysis*, vol. 54, pp. 1051-1065, 2010.

[90] B Walczak & DL Massart, "Noise suppression and signal compression using the wavelet packet transform," *Chemometrics and Intelligent Laboratory Systems*, vol. 36, pp. 81-94, 1997.

[91] K Jetter, U Depczynski, K Molt, A Niemöller, "Principles and applications of wavelet transformation to chemometrics," *Analytica Chimica Acta*, vol. 420, pp. 169-180, 2000.

[92] JC Cobas, PG Tahoces, M Martin-Pastor, M Penedo, FJ Sadina, "Wavelet-based ultra-high compression of multidimensional NMR data sets," *Journal of Magnetic Resonance*, vol. 168, pp. 288-295, 2004.

[93] RKH Galvão, MCU Araújo, TCB Saldanha, V Visani, MF Pimentel, "Estudo comparativo sobre filtragem de sinais instrumentais usando transformadas de Fourier e wavelet," *Químimica Nova*, vol. 24, pp. 874-884, 2001.

[94] M Misiti, Y Misiti, G Oppenheim, JM Poggi, *Wavelet toolbox for use with matlab*. Natick, USA: The MathWorks, 2002.

[95] MS Reis, PM Saraiva, BR Bakshi, "Denoising and signal-to-noise ratio enhancement: wavelet transform and Fourier transform," in *Comprehensive Chemometrics*, Stephen D Brown, Romà Tauler, and Beata Walczak, Eds. Amsterdam, The Netherlands: Elsevier B. V., 2009, ch. 2.03, pp. 25-54.

[96] BR Bakshi, "Multiscale PCA with application to multivariate statistical process monitoring," *American Institute of Chemical Engineers*, vol. 44, pp. 1596-1610, 44.

[97] TM Young, PM Winistorfer, S Wang, "Multivariate control charts of MDF and OSB vertical density profile attributes," *Forest Products Journal*, vol. 49, pp. 79-86, 1999.

[98] T Kourti, "Multivariate statistical process control and process control, using latent variables," in *Comprehensive Chemometrics*, Stephen D Brown, Romà Tauler, and Beata Walczak, Eds. Amsterdam, The Netherlands: Elsevier B.V., 2009, ch. 4.02, pp. 21-54.

[99] Y Hong, W ZhongRu, B TengFei, Z Lan, "Multivariate analysis in dam monitoring data with PCA," *Techonolgical Sciences*, vol. 53, pp. 1088-1097, 2010.

[100] GS Kapur, A Ecker, R Meusinger, "Establishing quantitative structure-property relationships (QSPR) of diesel samples by proton-NMR & multiple linear regression (MLR) analysis," *Energy & Fuels*, vol. 15, pp. 943-948, 2001.

[101] B Basu, GS Kapur, AS Sarpal, R Meusinger, "A neural network approach to the prediction of cetane number of diesel fuels using nuclear magnetic ressonance (NMR) spectroscopy," *Energy & Fuels*, vol. 17, pp. 1570-1575, 2003.

[102] FJ Hidalgo & R Zamora, "Edible oil analysis by high-resolution nuclear magnetic resonance spectroscopy: recent advances and future perspectives," *Trends in Food Science & Technology*, vol. 14, pp. 499-506, 2003.

[103] BR Cook, PJ Berlowitz, BG Silbernagel, DA Sysyn, "Use of ^{13}C NMR spectroscopy to produce optimum Fischer-Tropsch diesel fuels and blend stocks," United States Patent 6,210,559, 2003.

[104] AA Refaat, "Correlation between the chemical structure of biodiesel and its physical properties," *International Journal of Environmental Science and Technology*, vol. 6, pp. 677-694, 2009.

[105] JA Rodrigues Jr, FP Cardoso, ER Lachter, LRM Estevão, E Lima, RSV Nascimento, "Correlating Chemical Structure and Physical Properties of Vegetable Oil Esters," *Journal of the American Chemical Society*, vol. 83, pp. 353-357, 2006.

[106] J Boccard, A Kalousis, M Hilario, P Lantéri, M Hanafi, G Mazerolles, J-L Wolfender, P-A Carrupt, S Rudaz, "Standard machine learning algorithms applied to UPLC-TOF/MS metabolic fingerprinting for the discovery of wound biomarkers in Arabidopsis thaliana," *Chemometrics and Intelligent Laboratory Systems*, vol. 104, pp. 20–27, 2010.

[107] H-W Cho, SB Kim, MK Jeong, Y Park, TR Ziegler, DP Jones, "Genetic algorithm-based feature selection in high-resolution NMR spectra," *Metabolomics*, vol. 4, pp. 141–149, 2008.

[108] JK Nicholson, JC Lindon, E Holmes, "Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data,"

Xenobiotica, vol. 29, pp. 1181-1189, 1999.

[109] M Berk, T Ebbels, G Montana, "A statistical framework for biomarker discovery in metabolomic time course data," *Bioinformatics*, vol. 27, pp. 1979-1985, 2011.

[110] X Lin, Q Wang, P Yin, L Tang, Y Tan, H Li, K Yan, G Xu, "A method for handling metabonomics data from liquidchromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection," *Metabolomics*, vol. 7, pp. 549–558, 2011.

[111] J Sun, LK Schnackenberg, RD Holland, TC Schmitt, GH Cantor, YP Dragan, RD Beger, "Metabonomics evaluation of urine from rats given acute and chronic doses of acetaminophen using NMR and UPLC/MS," *Journal of Chromatography B*, vol. 871, pp. 328–340, 2008.

[112] JJ Jansen, HCJ Hoefsloot, J van der Greef, ME Timmerman, AK Smilde, "Multilevel component analysis of time-resolved metabolic fingerprinting data," *Analytica Chimica Acta*, vol. 530, pp. 173–183, 2005.

[113] JJ Jansen, E Szymanska, HCJ Hoefsloot, DM Jacobs, K Strassburg, AK Smilde, "Between Metabolite Relationships: an essential aspect of metabolic change," *Metabolomics*, vol. 8, pp. 422-432, 2012.

[114] W Weckwerth, ME Loureiro, K Wenzel, O Fiehn, "Differential metabolic networks unravel the effects of silent plant phenotypes," *PNAS*, vol. 101, pp. 7809–7814, 2004.

[115] Z Ramadan, D Jacobs, M Grigorov, S Kochhar, "Metabolic profiling using principal component analysis, discriminant partial least squares, and genetic algorithms," *Talanta*, vol. 68, pp. 1683–1691, 2008.

[116] JA Westerhuis, EJJ van Velzen, HCJ Hoefsloot, AK Smilde, "Multivariate paired data analysis: multilevel PLS-DA versus OPLS-DA," *Metabolomics*, vol. 6, pp. 119–128, 2010.

[117] P van den Brink & CJF Braak, "Principal response curves: analysis of time-dependent multivariate responses of biological community to stress," *Environmental Toxicology and Chemistry*, vol. 18, pp. 138-148, 1999.

[118] HC Keun, TMD Ebbels, ME Bollard, O Beckonert, H Antti, E Holmes, JC Lindon, JK Nicholson, "Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles," *Chemical Research in Toxicology*, vol. 17, pp. 579-587, 2004.

[119] N Yoshizuka, YY Chadani, V Krishnan, SL Zeichner, "Human immunodeficiency virus type 1 Vpr-dependent cell cycle arrest through a mitogen-activated protein kinase signal transduction pathway," *Journal of*

Virology, vol. 79, pp. 11366–11381, 2005.

[120] National Center for Biotechnology Information. Número de acesso: GSE2296. [Online]. "http://www.ncbi.nlm.nih.gov/geo/". Acesso em 05 de Maio de 2012, às 12h.

[121] L Stahle & S Wold, "Analysis of Variance (ANOVA)," *Chemometrics and Intelligent Laboratory Systems*, vol. 6, pp. 259-272, 1989.

[122] L Stahle & S Wold, "Multivariate Analysis of Variance (MANOVA)," *Chemometrics and Intelligent Laboratory Systems*, vol. 9, pp. 127-141, 1990.

[123] G Zwanenburg, HCJ Hoefsloot, JA Westerhuis, JJ Jansen, AK Smilde, "ANOVA–principal component analysis and ANOVA–simultaneous component analysis: a comparison," *Journal of Chemometrics*, vol. 25, pp. 561-567, 2011.

[124] AK Smilde, ME Timmerman, MMWB Hendriks, JJ Jansen, HCJ Hoefsloot, "Generic framework for high-dimensional fixed-effects ANOVA," *Briefings in Bioinformatics*, vol. 13, pp. 524-535, 2012.

[125] HCJ Hoefsloot, D Vis, JA Westerhuis, AK Smilde, JJ Jansen, "Multiset data analysis: ANOVA simultaneous component analysis and related methods," in *Comprehensive Chemometrics*, Stephen D Brown, Romà Tauler, and Beata Walczak, Eds. Amsterdam, The Netherlands: Elsevier B. V., 2009, ch. 2.23, pp. 453-472.

[126] AK Smilde, HCJ Hoefsloot, JA Westerhuis, "The geometry of ASCA," *Journal of Chemometrics*, vol. 22, pp. 464-471, 2008.

[127] AK Smilde, JJ Jansen, HCJ Hoefsloot, RJAN Lamers, J van der Greef, ME Timmerman, "ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data," *Bioinformatics*, vol. 21, pp. 3043–3048, 2005.

[128] JJ Jansen, HCJ Hoefsloot, J van der Greef, ME Timmerman, JA Westerhuis, AK Smilde, "ASCA: analysis of multivariate data obtained from an experimental design," *Journal of Chemometrics*, vol. 19, pp. 469-481, 2005.

[129] MPH Verouden, JA Westerhuis, MJ van der Werf, AK Smilde, "Exploring the analysis of structured metabolomics data," *Chemometrics and Intelligent Laboratory Systems*, vol. 98, pp. 88–96, 2009.

[130] PB Harrington, NE Vieira, J Espinoza, JK Nien, R Romero, AL Yergey, "Analysis of variance–principal component analysis: A soft tool for proteomic discovery," *Analytica Chimica Acta*, vol. 544, pp. 118-127, 2005.

[131] NT Trendafilov, "Orthonormality-constrained INDSCAL with non-negative saliences," *Lecture Notes in Computer Science*, vol. 3044, pp. 952-960, 2004.

[132] Y Takane, K Jung, H Hwang, "An acceleration method for ten Berge et al.'s algorithm for orthogonal INDSCAL," *Computational Statistics*, vol. 25, pp. 409-428, 2010.

[133] JMF ten Berge & HAL Kiers, "Some clarifications of the candecomp algorithm applied to INDSCAL," *Psychometrika*, vol. 56, pp. 317-326, 1991.

[134] A Stegeman, "On uniqueness conditions for Candecomp/Parafac and Indscal with full column rank in one mode," *Linear Algebra and its Applications*, vol. 431, pp. 211-227, 2009.

[135] JJ Jansen, E Szymanska, HCJ Hoefsloot, AK Smilde, "Individual differences in metabolomics: individualized responses and between-metabolite relationships," *Metabolomics*, vol. 8, pp. S94-S104, 2012.

[136] R Bro, E Acar, T Kolda, "Resolving the sign ambiguity in the singular value decomposition," SANDIA National Laboratories, Livermore, SAND2007-6422, 2007.

[137] Swiss Institute of Bioinformatics & Geneva Bioinformatics SA. NextProt Beta: Exploring the universe of human proteins. [Online]. "http://www.nextprot.org/". Acesso em 22 de Abril de 2013, às 14h.

[138] F Wu, J Garcia, D Sigman, R Gaynor, "Tat regulates binding of the human immunodeficiency virus trans-activating region RNA loop-binding protein TRP-185," *Genes & Development*, vol. 5, pp. 2128-2140, 1991.

[139] A Trkola, "HIV–host interactions: vital to the virus and key to its inhibition," *Current Opinion in Microbiology*, vol. 7, pp. 407-411, 2004.

[140] A Zaheer & R Lim, "In vitro inhibition of MAP kinase (ERK1/ERK2) activity by phosphorylated glia maturation factor (GMF)," *Biochemistry*, vol. 35, pp. 6283-6288, 1996.

[141] R Lim & A Zaheer, "In vitro enhancement of p38 mitogen-activated protein kinase activity by phosphorylated glia maturation factor," *The Journal of Biological Chemistry*, vol. 271, pp. 22953-22956, 1996.

[142] YW Heng & CG Koh, "Actin cytoskeleton dynamics and the cell division cycle," *The International Journal of Biochemistry & Cell Biology*, vol. 42, pp. 1622-1633, 2010.

[143] Weizmann Institute of Science. Genecards: The human gene compendium. [Online]. "http://www.genecards.org/". Acesso em 22 de Abril de 2013, às 14h.

[144] G Heller, B Ziegler, A Brandstetter, S Novak, M Rudas, G Hennig, M Gehrmann, T Acht, S Zöchbauer-Müller, M Filipits, "CDK10 is not a target for aberrant DNA methylation in breast cancer," *Anticancer Research*, vol. 29, pp. 3939-3944, 2009.

[145] EI Christensen & H Birn, "Megalin and cubilin: multifunctional endocytic receptors," *Nature Reviews Molecular Cell Biology,* vol. 3, pp. 258-268, 2002.

[146] JL Andersen & V Planelles, "The role of Vpr in HIV-1 pathogenesis," *Current HIV Research*, vol. 3, pp. 43-51, 2005.