



UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

TESE DE DOUTORADO

***UMA ESTRATÉGIA DE SELEÇÃO DE VARIÁVEIS BASEADA NA
OTIMIZAÇÃO POR COLÔNIA DE FORMIGAS EM
ANÁLISE DISCRIMINANTE LINEAR***

Aline Santos de Pontes

João Pessoa - PB
Dezembro de 2020



UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE QUÍMICA
PROGRAMA DE PÓS-GRADUAÇÃO EM QUÍMICA

TESE DE DOUTORADO

***UMA ESTRATÉGIA DE SELEÇÃO DE VARIÁVEIS BASEADA NA
OTIMIZAÇÃO POR COLÔNIA DE FORMIGAS EM
ANÁLISE DISCRIMINANTE LINEAR***

Aline Santos de Pontes*

Tese apresentada ao Programa de Pós-Graduação em Química/CCEN/UFPB como parte dos requisitos para obtenção do título de Doutora em Química, área de concentração Química Analítica.

Orientador: Prof. Dr. Edvan Cirino da Silva (UFPB)

***Bolsista do Conselho Nacional de desenvolvimento Científico e tecnológico**

João Pessoa - PB
Dezembro de 2020

Catálogo na publicação
Seção de Catalogação e Classificação

P814e Pontes, Aline Santos de.

Uma estratégia de seleção de variáveis baseada na
otimização por colônia de formigas em análise
discriminante linear / Aline Santos de Pontes. - João
Pessoa, 2020.
163 f. : il.

Orientação: Edvan Cirino da Silva.
Tese (Doutorado) - UFPB/CCEN.

1. Química analítica. 2. Seleção de variáveis. 3.
Colônia de formigas. 4. Análise discriminante. I.
Silva, Edvan Cirino da. II. Título.

UFPB/BC

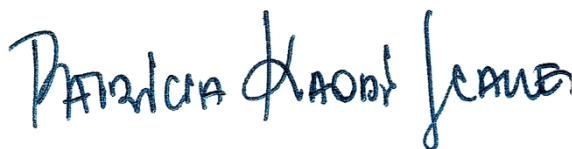
CDU 543(043)

Uma estratégia de seleção de variáveis baseada na otimização por colônia de formigas em análise discriminante linear

Tese de Doutorado apresentada pela aluna Aline Santos de Pontes e aprovada pela banca examinadora em 14 de dezembro de 2020.



Prof. Dr. Edvan Cirino da Silva
Orientador/Presidente



Profa. Dra. Patrícia Kaori Soares
Examinadora



Profa. Dr. Kássio Michell Gomes de Lima
Examinador



Prof. Dr. Sherlan Guimarães Lemos
Examinador



Prof. Dr. Márcio José Coelho de Pontes
Examinador

*Tu, Senhor, guardarás em perfeita paz aquele cujo propósito está firme, porque em ti confia. **Isaías 26:3***

A trindade (Deus, Jesus e Espírito Santo), pela vida, salvação e amor incondicional, dedico...

A minha família, por sempre terem se dedicado, apoiado e ajudado ao máximo, dedico...

AGRADECIMENTOS

- ✚ Ao meu amado Pai Celestial, pelo discernimento e força para concluir a tese de doutorado;
- ✚ Aos meus pais Adilson Dias de Pontes e Maria de Fátima Santos pelo apoio, carinho, cuidado, amor e incentivo em todos os momentos;
- ✚ Aos meus irmãos, Adilson Filho, Thereza Klein, Wellando e Cristiane, pelo carinho e amor;
- ✚ Ao meu esposo, Arian de Souza Santana, pelas orações, amor, companheirismo, dedicação e apoio nos momentos difíceis;
- ✚ Aos meus cunhados e cunhadas pelo carinho, incentivos e amor;
- ✚ Ao professor Edvan Cirino da Silva, pela excelente orientação e confiança depositada durante toda iniciação científica, mestrado e doutorado.
- ✚ Ao professor Mário César Ugulino de Araújo, por sempre acreditar nos seus alunos e não medir esforços para ajudar no desenvolvimento acadêmico. Obrigada pela grande contribuição, ensinamentos, confiança e sugestões;
- ✚ Ao amigo e hoje professor Adriano de Araújo Gomes, por ter me ajudado desde a elaboração do projeto do doutorado até o desenvolvimento e conclusão da tese, o meu muito obrigado por tudo;
- ✚ Aos professores do PPGQ/UFPB pelos ensinamentos durante as disciplinas cursadas;
- ✚ Aos amigos que vão além do lado profissional e fazem parte da minha vida;
- ✚ Aos amigos que ganhei na UFPB e no LAQA durante todo esse tempo;
- ✚ À EMBRAPA Arroz e Feijão, sediada em Santo Antônio de Goiás-GO, por ter cedido, para o trabalho realizado nessa Tese no âmbito de Acordo de Transferência de Material firmado com a UFPB, os cultivares de feijões, incluindo o transgênico;
- ✚ Ao PPGQ/UFPB e a CAPES pela bolsa concedida.

SUMÁRIO

LISTA DE FIGURAS	iv
LISTA DE TABELAS	ix
LISTA DE ABREVIATURAS	x
RESUMO.....	xi
ABSTRACT	xii

CAPÍTULO I

INTRODUÇÃO

1. INTRODUÇÃO.....	2
1.1 Caracterização geral da problemática	2
1.2 Objetivos	4
1.2.1 Objetivos gerais	4
1.2.2 Objetivos específicos	5

CAPÍTULO II

FUNDAMENTAÇÃO TEÓRICA

2. FUNDAMENTAÇÃO TEÓRICA	7
2.1 Dados em química analítica	7
2.1.1 Métodos de Análise Multivariada.....	7
2.2.2 Pré-Processamento dos dados.....	9
2.2.3 Reconhecimento de Padrões	11
2.2.3.1 Técnicas de Reconhecimento de Padrões não Supervisionadas ...	12
2.2.3.1.1 Análise po componebtes Principais-PCA	12
2.2.3.2 Técnicas de Reconhecimento de Padrões Supervisionadas	14
2.2.3.2.1 Análise Discriminante pelos Mínimos-Quadrados Parciais-PLS-DA	15
2.2.3.2.2 Análise Discriminante Linear-LDA	17
2.3. Seleção de variáveis.....	19
2.3.1 Algoritmo das Projeções Sucessivas.....	21
2.3.2 Algoritmo genético.	23
2.3.3 Colônias de Formigas.....	28
2.3.3.1 Colônias de Formigas reais	28
2.3.3.2 Colônias de Formigas artificiais	31
2.3.4 Algoritmo de otimização por colônia de formigas	32

2.3.4.1 Deposição e evaporação dos feromônios.....	34
2.3.4.2 Atualização da colônia de formigas	35
2.3.5 Aplicações da Técnica de Otimização de Colônia de Formigas no campo da química	36
2.4 Parâmetros de validação dos modelos de classificação	39

CAPÍTULO III

ALGORITMO COLÔNIA DE FORMIGAS PARA CLASSIFICAÇÃO

3. Descrição do ACO-LDA proposto para seleção de variáveis	43
3.1 Codificação das formigas.....	44
3.2. Geração da primeira colônia de formigas e do vetor feromônio	45
3.3. Avaliação de aptidão	46
3.4. Atualização da trilha de feromônio	48
3.5. Atualização da colônia	50
3.6. Critério de parada	51

CAPÍTULO IV

ESTUDO DE CASO I E II: CLASSIFICAÇÃO DE ÓLEOS VEGETAIS E CHÁS

4. Classificação de óleos vegetais e chás	55
4.1. Introdução	55
4.1.1. Óleos vegetais comestíveis	55
4.1.2. Chás.....	56
4.2. Metodologia	57
4.2.1. Conjunto de dados para análise multivariada	57
4.2.2. Procedimentos quimiométricos.....	58
4.3. Resultados e discussões.....	60
4.3.1. Classificação do óleo comestível em relação ao tipo de oleaginosa.....	60
4.3.1.1. Perfil espectral das amostras de óleo comestível	60
4.3.1.2 ACO-LDA.....	62
4.3.1.3. GA-LDA.....	66
4.3.1.4. PLS-DA	68
4.3.2. Classificação simultânea de chá com respeito a variedade e origens geográficas.....	72
4.3.2.1. Perfil espectral das amostras de chás.....	72
4.3.2.2. Classificação	74
4.3.2.2.1. ACO-LDA	74

4.3.2.2.2. GA-LDA	77
4.3.2.2.3. PLS-DA.....	79

CAPÍTULO V

CLASSIFICAÇÃO DE TRÊS CULTIVARES DE FEIJÃO

5. Classificação de três cultivares de feijão	84
5.1. Introdução	84
5.1.1. Feijão	84
5.1.1.1. Cultivar BRS Estilo.....	86
5.1.1.2. Cultivar BRS Pontal.....	86
5.1.1.3. Cultivar BRSFC401RMD	87
5.1.2. Composição Química.....	87
5.1.3. Pragas na cultura do feijoeiro	88
5.1.4. Mosaico dourado	89
5.1.5. Os organismos geneticamente modificados ou transgênicos.....	92
5.1.6. O feijão geneticamente modificado	93
5.1.7. Ferramentas na detecção dos feijões convencionais e transgênicos	95
5.2. Experimental	96
5.2.1. Amostras	96
5.2.2 Aquisições dos Espectros reflectância na região do NIR	97
5.2.3. Análise Quimiométrica	100
5.3. Resultados e Discussão.....	102
5.3.1. Perfil espectral das amostras de feijão	102
5.3.2. Análise por Componentes Principais (PCA).....	105
5.3.3. Métodos de Reconhecimento de Padrões Supervisionados	109
5.3.3.1 ACO-LDA.....	110
5.3.3.2. SPA-LDA	112
5.3.3.3. GA-LDA.....	115
5.3.3.4. PLS-DA	117

CAPÍTULO VI

CONCLUSÕES E PERSPECTIVAS FUTURAS

6.1 Conclusões	122
6.2 Perspectivas futuras.....	123

REFERÊNCIAS	124
--------------------------	-----

APÊNDICE	145
-----------------------	-----

LISTA DE FIGURAS

Figura 2.1. PC1 e PC2 obtidas a partir da combinação linear das variáveis originais. As linhas tracejadas na cor verde correspondem aos escores e as linhas tracejadas na cor pretas representam os pesos.....	13
Figura 2.2. Organização dos dados para análise multivariada.....	14
Figura 2.3. Representação do modelo preditivo PLS-DA.....	16
Figura 2.4. Representação esquemática da análise discriminante linear.....	18
Figura 2.5. Codificação binária usada na seleção de variáveis do GA.....	24
Figura 2.6. Representação dos cromossomos gerados de maneira aleatória.....	25
Figura 2.7. Esquema de processo de cruzamento no algoritmo genético..	26
Figura 2.8. Esquema de processo de mutação no algoritmo genético.....	27
Figura 2.9. Formigas na busca do melhor caminho até o alimento. a) as formigas forrageiam aleatoriamente. b) formação da trilha de feromônio. c) trilha de feromônio formada.....	28
Figura 2.10. Formigas na busca do melhor caminho até o alimento. a) as formigas são deixadas livres para escolher o caminho. b) as formigas convergem para um dos caminhos probabilidade iguais.....	29
Figura 2.11. Comportamento das formigas que convergem para a ponte mais curta.....	30
Figura 3.1. Codificação de como as variáveis são selecionadas por colônia de formigas.....	44
Figura 3.2. Inicialização da primeira colônia pelo processo randômico...	45
Figura 3.3. Avaliação do poder discriminante para cada modelo construído das formiga iniciais.....	48
Figura 3.4. Atualização da trilha de feromônio.....	49
Figura 3.5. Atualização da colônia.....	51
Figura 3.6. Fluxograma do algoritmo ACO-LDA. A. Sem Double Monte Carlo e B. Com Double Monte Carlo.....	52

Figura 4.1. Estruturas moleculares dos ácidos (a) oleico, (b) linoléico, (c) linolênico.....	56
Figura 4.2. Espectros UV-VIS das 4 classes das amostras de óleos vegetais comestíveis analisados.....	60
Figura 4.3. Espectros médios de canola (linha azul), soja (linha verde), milho (linha amarela) e girassol (linha vermelha).....	61
Figura 4.4. Espectros UV-VIS de todas as 4 categorias das amostras de óleos vegetais comestíveis analisados.....	62
Figura.4.5. Comprimento de onda ou variáveis latentes (círculos brancos) selecionados pelos algoritmos ACO-LDA dos dados do UV-Vis.....	63
Figura 4.6. Valores VIF (linha azul sólida) para as variáveis selecionadas (círculos brancos com borda preta) por ACO-LDA dos dados do UV-Vis....	64
Figura. 4.7. Gráficos de escores e mapa térmico dos pesos do modelo ACO-LDA (A e B), para classificação de óleos comestíveis. Círculo azul (canola), triângulo vermelho (soja), quadrado amarelo (milho) e diamante verde (girassol).....	65
Figura 4.8: Comprimento de onda ou variáveis latentes (círculos brancos) selecionados pelos algoritmos GA-LDA dos dados do UV-Vis.....	66
Figura 4.9. Valores VIF (linha azul sólida) para as variáveis selecionadas (círculos brancos com borda preta) por GA-LDA dos dados do UV-Vis.....	66
Figura 4.10. Gráficos de escores e mapa térmico dos pesos do modelo GA-LDA, para classificação de óleos comestíveis. Círculo azul (canola), triângulo vermelho (soja), quadrado amarelo (milho) e diamante verde (girassol)..	67
Figura 4.11. Número de variáveis latentes selecionados pelo PLS-DA para classificação de óleos comestíveis.....	68
Figura 4.12. Grafico de escores (A) e pesos (B) do PLS-DA. Círculo azul (canola), triângulo vermelho (soja), quadrado amarelo (milho) e diamante verde (girassol).....	69
Figura 4.13. Espectros bruto NIR de todas as 5 categorias das amostras de chás analisados.....	72
Figura 4.14.. Espectros NIR das 5 classes das amostras de chás analisados.....	73
Figura 4.15. Espectros médios de chá de AG (linha azul), AB (linha verde), BG (linha vermelha), BB (linha ciano) e SB (linha magenta).....	74

Figura 4.16. Comprimento de onda ou variáveis latentes (círculos brancos) selecionados pelos algoritmos ACO-LDA nos espectros NIR.....	75
Figura 4.17. Valores VIF (linha azul sólida) para as variáveis selecionadas (círculos brancos com borda preta) por ACO-LDA dos dados do Nir.....	76
Figura 4.18. Gráfico de escores e mapa de calor de pesos dos modelos ACO-LDA (A e B), para classificação de chás. AG (círculo azul), AB (quadrada verde), BG (triângulo vermelho), BB (diamante ciano) e SB (triângulo margenta de cabeça para baixo).....	76
Figura 4.19. Comprimento de onda ou variáveis latentes (círculos brancos) selecionados pelos algoritmos GA-LDA nos espectros NIR.....	78
Figura 4.20. Valores VIF (linha azul sólida) para as variáveis selecionadas (círculos brancos com borda preta) por GA-LDA dos dados do Nir.....	78
Figura 4.21. Gráfico de escores e mapa de calor de pesos dos modelos GA-LDA (A e B), para classificação de chás. AG (círculo azul), AB (quadrada verde), BG (triângulo vermelho), BB (diamante ciano) e SB (triângulo margenta de cabeça para baixo).....	79
Figura 4.22. Número de variáveis latentes selecionados pelo PLS-DA para classificação de chás.....	80
Figura 4.23: Gráfico de escores (A) e pesos (B) dos modelos ACO-LDA (A e B), para classificação de chás. AG (círculo azul), AB (quadrada verde), BG (triângulo vermelho), BB (diamante ciano) e SB (triângulo margenta de cabeça para baixo).....	80
Figura 5.1 Agente transmissor da doença do mosaico dourado (mosca branca).....	90
Figura 5.2. Cultivar de feijão afetado pelo mosaico dourado.....	91
Figura 5.3. Planta geneticamente modificada (esquerda), imune à doença do mosaico dourado e uma planta não modificada (direita).....	95
Figura 5.4. Separação das sementes das diferentes cultivares de feijões.....	97
Figura 5.5. espectrômetro portátil DLP NIRscan Nano EVM.....	98
Figura 5.6. Programa NIRscan Nano GUI v1.1.9.....	98
Figura 5.7. (a) Célula de <i>Teflon</i>	99

Figura 5.8. Padrão de reflectância.....	100
Figura 5.9. Espectros originais na região do infravermelho próximo das sementes dos feijões BRSFC401RMD (verde), BRSPontal (amarelo) e BRSEstilo (azul).....	102
Figura 5.10. Gráfico dos escores PC1 x PC2 para as sementes de feijões na região do NIR. (Verde: BRSFC401RMD; Amarelo: BRSPontal; Azul: BRSEstilo).....	103
Figura 5.11 Espectros pré-processados com 1º derivada utilizando o algoritmo S.G com polinômio e janela de: a) 11 pontos e b) 21 pontos; 2º derivada utilizando o algoritmo S.G com polinômio e janela de c) 11 pontos e d) 21 pontos; e) MSC; f) SNV; g) Baseline. (Verde: BRSFC401RMD; Amarelo: BRSPontal; Azul: BRSEstilo).....	104
Figura 5.12. PCA para as sementes de feijões na região do NIR na faixa de 900 a 1700. a) 1º Savitzky-Golay (11); b) 1º Savitzky-Golay (21); c) 2º Savitzky-Golay (11); d) 2º Savitzky-Golay (21) e) MSC; f) SNV; g) Baseline. (Verde: BRSFC401RMD; Amarelo: BRSPontal; Azul: BRSEstilo).....	106
Figura 5.13. PCA para as sementes de feijões na região do NIR de 900 a 1630. a) 1º Savitzky-Golay (11); b) 1º Savitzky-Golay (21); c) 2º Savitzky-Golay (11); d) 2º Savitzky-Golay (21) e) MSC; f) SNV; g) Baseline. (Verde: BRSFC401RMD; Amarelo: BRSPontal; Azul: BRSEstilo).....	107
Figura 5.14. Gráficos dos pesos <i>versus</i> variáveis de PC1(azul) e PC2(vermelho) de todas as classes.....	108
Figura 5.15. Espectro médio para cada tipo semente analisado. Os onze comprimentos de ondas selecionados pelo ACO-LDA encontram-se indicados em círculos.....	110
Figura 5.16. Gráfico dos escores da função discriminante 1 para a espectroscopia NIR das sementes BRSFC401RMD (●), BRSPontal (●) e BRSEstilo (●) obtidos pela ACO-LDA.....	111
Figura 5.17. Gráfico dos pesos das variáveis selecionadas pelo ACO-LDA para as três cultivares dos feijões.....	112
Figura 5.18. Espectro médio com as variáveis selecionadas pelo SPA..	113

- Figura 5.19.** Gráfico dos escores da função discriminante 1 para a espectroscopia NIR das sementes BRSFC401RMD (●), BRSPontal (●) e BRSRstilo (●) obtidos pela SPA-LDA.....113
- Figura 5.20.** Gráfico dos pesos das variáveis selecionadas pelo SPA-LDA para as três cultivares dos feijões.....114
- Figura 5.21.** Espectro médio para cada tipo semente analisado. (a)transgênicos, (b) convencionais. Os seis comprimentos de ondas selecionados pelo GA-LDA encontram-se indicados em círculos.....115
- Figura 5.22.** Gráfico dos escores da função discriminante 1 para a espectroscopia NIR das sementes BRSFC401RMD (●), BRSPontal (●) e BRSRstilo (●) obtidos pela GA-LDA.....116
- Figura 5.23.** Gráfico dos pesos das variáveis selecionadas pelo SPA-LDA para as três cultivares dos feijões.....117
- Figura 5.24:** Taxa de erro versus o número de variáveis latentes incluídas no modelo.....118

LISTA DE TABELAS

Tabela 2.1: Exemplo de uma matriz de confusão.....	40
Tabela 4.1: Parâmetros de adequação e validação dos modelos PLS-DA, GA-LDA e ACO-LDA empregados para a classificação de óleos vegetais comestíveis.....	70
Tabela 4.2: Matriz confusão de predição dos óleos comestíveis.....	71
Tabela 4.3: Parâmetros de adaptação e validação dos modelos PLS-DA, GA-LDA e ACO-LDA para classificação de chás.....	81
Tabela 4.4: Matriz de confusão para previsão de chás em um conjunto de teste independente.....	82
Tabela 5.1: Principais doenças que afetam o feijoeiro.....	89
Tabela 5.2: Número de amostras de treinamento, validação e teste para as classes de sementes de feijões.....	101
Tabela 5.3: Taxa de classificação correta obtida pelos modelos de classificação LDA-ACO, LDA-GA e LDA-SPA aplicados aos dados do NIR para o conjunto de teste. Os valores em parênteses indicam o número de variáveis ou fatores selecionados.....	109
Tabela 5.4: Resultados de previsão para um conjunto de treinamento: matriz de confusão, sensibilidade e especificidade.....	119
Tabela 5.5: Resultados de previsão para um conjunto de teste: matriz de confusão, sensibilidade e especificidade.....	120

LISTA DE ABREVIATURAS

AC - Colônia de Formigas, do inglês Ant Colony.

ACO - Algoritmo de Colônia de Formigas.

BGMV - Vírus do mosaico dourado do feijão, do inglês Bean golden mosaic virus.

FD - Funções Discriminantes.

GA - Algoritmo Genético, do inglês Genetic Algorithm.

HCA - Análise de Agrupamentos Hierárquicos, do inglês Hierarchical Cluster Analysis.

ISAAA - Serviço Internacional para a Aquisição de Aplicações em Agrobiotecnologia.

LDA - Análise Discriminante Linear, do inglês Linear Discriminant Analysis.

MLR - Regressão Linear Múltipla, do inglês Multiple Linear Regression.

MSC - Correção de Sinal Múltiplo, do inglês multiplicative scatter correction.

NIR - Infravermelho Próximo, do inglês Near Infra Red.

OGMs - Organismos geneticamente modificados.

PGMs - Plantas geneticamente modificadas.

PC - Componentes Principais, do inglês Principal Components.

PCA - Análise por Componentes Principais, do inglês Principal Component Analysis

PCR - Regressão por Componentes Principais, do inglês Principal Component Regression.

PLS - Mínimos quadrados parciais, do inglês Partial Least Squares

RMSECV - Raiz Quadrada do Erro Médio Quadrático de Validação Cruzada, do inglês Root Mean Square Error of Cross-Validation

RMSEV - Raiz Quadrada do Erro Médio Quadrático de Validação, do inglês Root Mean Square Error of Validation.

RMD - Resistência ao mosaico-dourado.

RP - Reconhecimento de Padrão, do inglês Pattern Recognition.

SPA - Algoritmo das projeções sucessivas, do inglês Successive Projections Algorithm.

SNV - Variação normal padrão, do inglês standart normal variate.

RESUMO

Titulo: *"Uma estratégia de seleção de variáveis baseada na otimização por colônia de formigas em análise discriminante linear"*

Neste trabalho é apresentado o uso de uma nova estratégia de otimização bioinspirada em Colônia de Formigas para seleção de variáveis em problemas de classificação via Análise Discriminante Linear (LDA). O algoritmo proposto, denominado ACO-LDA, foi implementado em ambiente MatLab (como também todos os demais cálculos quimiométricos) buscando minimizar uma função de custo associada ao risco médio de erro de classificação G_{cost} . O ACO-LDA foi avaliado usando dados espectrométricos de infravermelho próximo (NIR) e de ultravioleta visível (UV-Vis). No primeiro estudo, a espectrometria UV-VIS é adotada para classificar quatro tipos de óleos vegetais comestíveis (milho, soja, canola e girassol). No segundo caso, a espectrometria NIR é usada para discriminar amostras de chás com respeito às suas variedades (preto e verde) e origens geográficas (argentino, brasileiro e srilanquense). No terceiro caso, é empregado dados de espectrometria NIR em sementes de feijões segundo o tipo de cultivar: (perola, pontal e transgênico). Os resultados obtidos foram comparados a outros métodos de seleção de variáveis bem estabelecidos na literatura como Algoritmo genético (GA), o Algoritmo das Projeções Sucessivas (SPA), ambos associados às modelagens LDA e por Análise discriminante por mínimos quadrados parciais (PLS-DA). Nas duas primeiras aplicações, a taxa de classificação correta de previsão foram: 100%, 95,7% e 100% dos modelos ACO-LDA, GA-LDA e PLS-DA respectivamente para classificação das amostras de óleos vegetais (UV-Vis). Enquanto para classificação das amostras de chás (NIR), a taxa de classificação correta foram de 92%, 84% e 100% dos modelos ACO-LDA, GA-LDA e PLS-DA respectivamente. Já na terceira aplicação, a taxa de classificação correta foi de 100% para os quatro modelos (ACO-LDA, GA-LDA, SPA-LDA e PLS-DA) na classificação das amostras de feijões. Dessa forma, os resultados indicam que o ACO-LDA é promissor para classificação de amostras a partir de dados espectrométricos NIR e UV-Vis, possibilitando fornecer resultados compatíveis com aqueles obtidos por ferramentas quimiométricas poderosas e versáteis.

Palavras-chave: Seleção de variáveis, colônia de formigas, análise discriminante linear.

ABSTRACT

Title: "A variable selection strategy based on ant colony optimization in linear discriminant analysis"

In this work the use of a new biophysical optimization strategy in Ants' Colony for selection of variables in classification problems via Linear Discriminant Analysis (LDA) is presented. The proposed algorithm, called ACO-LDA, was implemented in MatLab environment (as well as all other chemometric calculations) and evaluated in two databases already studied and a database developed. In the first study, UV-VIS spectrometry was used to classify four types of edible vegetable oils (corn, soybean, canola and sunflower). In the second case, NIR spectrometry is used to discriminate samples of teas with respect to their varieties (black and green) and geographical origins (Argentinean, Brazilian, and Sri Lankan). In the third case, NIR spectrometry data were used in beans seeds according to the type of cultivar: (perola, pontal and transgenic). The results obtained were compared to other methods of selection of variables good established in the literature as Genetic Algorithm (GA), the Algorithm of Successive Projections (SPA), both associated to LDA modeling and the Discriminante Analysis by Partial Least Squares (PLS-DA). In the first two applications, a correct prediction classification rate was: 100%, 95.7% and 100% of the ACO-LDA, GA-LDA and PLS-DA models respectively for classifying the properties of vegetable plants (UV-Vis). While for the classification of teas (NIR), a correct classification rate was 92%, 84% and 100% of the ACO-LDA, GA-LDA and PLS-DA models respectively. In the third application, a correct classification rate was 100% for the four models (ACO-LDA, GA-LDA, SPA-LDA and PLS-DA) in the classification of debts. Thus, the results indicated that ACO-LDA is promising to classify from NIR and UV-Vis spectrometric data, making it possible to provide results compatible with those chosen by powerful and versatile chemometric tools.

Keywords: Variable selection, ant colony, linear discriminant analysis.

CAPÍTULO I

INTRODUÇÃO



1. INTRODUÇÃO

1.1 Caracterização geral da problemática

O avanço tecnológico permitiu o desenvolvimento de instrumentos analíticos modernos capazes de produzir uma grande quantidade de informações analíticas por amostra em curtos intervalos de tempo. Consequentemente, os dados gerados apresentam uma alta complexidade decorrente das muitas variáveis usualmente envolvidas nas medidas instrumentais. Desse modo, a aquisição e construção de modelos multivariados (de calibração ou classificação) tornaram-se uma prática usual em laboratórios de pesquisa em Química Analítica.

Na literatura são reportados muitos avanços no campo da instrumentação que permitem registros de sinais em múltiplos canais analíticos (comprimentos de onda, distribuição de frequências, potenciais, tempos de eluição)¹⁻⁴. Isso possibilita o desenvolvimento de métodos que necessitam de menor consumo de reagentes/amostras e de procedimentos não destrutivos e não invasivos⁵⁻⁶. Esses métodos costumam apresentar vantagens como rapidez, baixo custo, pouco ou nenhum tratamento prévio das amostras⁶⁻⁷. Para isso, usam-se técnicas espectroanalíticas (espectrofotometrias UV-Vis⁸ e NIR⁹ e fluorescência¹⁰, por exemplo), voltamétricas¹¹, etc.

Apesar desses avanços, a obtenção de um número elevado de dados pode envolver variáveis não informativas ou redundantes, necessitando assim de uma interpretação criteriosa dos resultados ¹². Para extrair o máximo de informação útil dos dados para abordar o problema focalizado, pode-se usar ferramentas matemáticas e estatísticas descritas na Quimiometria.

Em problemas de classificação, são usualmente realizadas medidas de propriedades (sinais) em diversas variáveis para um grupo de amostras representativas (objetos) de classe(s) conhecida(s) gerando um conjunto de dados multivariados. A estratégia de compressão bem estabelecida por análise de componentes principais (PCA) ou por mínimos quadrados parciais

(PLS) tem sido usada para fins de seleção de variáveis. Essas técnicas quimiométricas são ferramentas poderosas para lidar com alta dimensionalidade, multicolinearidade e ruído presentes em dados espectrométricos típicos ^{13,14}. No entanto, é necessário mais esforço para interpretar os resultados das predições devido à transformação matemática realizada no domínio de dados em escores, pesos e coeficientes de regressão, que são de interpretação química não trivial¹⁵. Uma alternativa importante à PCA e PLS é a análise discriminante linear (Linear Discriminant Analysis-LDA) que tem a grande vantagem de operar no domínio original dos dados, permitindo uma interpretação mais fácil dos resultados. No entanto, essa técnica necessita de seleção de variáveis para redução dimensionalidade do espaço e multicolinearidade entre variáveis ¹⁶.

O maior desafio dessas técnicas é selecionar variáveis mais informativas e não redundantes para a construção de modelos mais eficientes que os obtidos dos espectros completos. Pode-se ainda contemplar o princípio da parcimônia¹⁷ ao se obter modelos com desempenhos similares, adotando-se o menor número de variáveis selecionadas. Tais ferramentas de seleção buscam selecionar as variáveis minimamente multicolineares que contenham as informações relacionadas às características distintivas de cada classe. Assim, a modelagem pode ser realizada com base nessas variáveis a fim de construir modelos mais simples, robustos, eficientes e fáceis de interpretar.

As abordagens de problemas analíticos, baseadas em seleção de variáveis, têm sido amplamente utilizadas na calibração multivariada. Todavia, poucos trabalhos têm sido encontrados na literatura envolvendo o uso dessas ferramentas em problemas de classificação. Diversos algoritmos, reportados na literatura para seleção de variáveis em modelagem de classificação, têm sido desenvolvidos. A título de exemplos, podem ser mencionados: algoritmo genético, (GA, do inglês "Genetic Algorithm")¹⁸, stepwise (SW)¹⁹, o algoritmo das projeções sucessivas (SPA, do inglês Successive Projections Algorithm)^{20,21}, etc. Alguns trabalhos relatam o uso de algoritmos de seleção de variáveis acopladas ao LDA, tais

como SPA-LDA, GA-LDA e SW-LDA, os quais operam no domínio original dos dados. Estes produzem modelos com interpretação química mais fácil e desempenho similar aos obtidos pelos algoritmos PCA e PLS – que usam estratégias de compressão de dados.

Nas últimas décadas um novo campo de pesquisa, denominado bioinspiração, tem afluído e sido explorado nas áreas de engenharia e biomedicina^{22,23}. A bioinspiração consiste na busca da compreensão dos mecanismos e mimetização do comportamento de espécies animais (pássaros, peixes, etc)²⁴ ou mesmo vegetais (ervas daninhas²⁵, por exemplo). A bioinspiração tem como finalidade aplicar os algoritmos, elaborados para a mimetização de interesse, à resolução de problemas e/ou desenvolvimento de novas tecnologias. A título de exemplo, destaca-se a otimização por colônia de formigas (Ant Colony Optimization-ACO), cujo algoritmo foi proposto por Dorigo et al.²⁶ e aplicado pela primeira vez no problema do caixeiro viajante (TSP). As aplicações e eficácia de vários algoritmos baseados no ACO, na resolução de problemas complexos de otimização, tem sido demonstrado e reportado na literatura ^{27 -29}.

Neste trabalho, propõe-se o uso do ACO para desenvolver um novo algoritmo para seleção de variáveis em modelagem multivariada para fins de classificação via LDA. Esta proposição se justifica pelo fato do ACO ter sido reportado na literatura apenas no contexto da calibração multivariada, sendo acoplado à técnica PLS³⁰.

1.2 Objetivos

1.2.1 Objetivo geral

O objetivo central deste trabalho foi propor um algoritmo, baseado na otimização por colônia de formigas, para seleção de variáveis em modelagem multivariada envolvendo problemas de classificação.

1.2.2 Objetivos específicos

- Introduzir a função custo apropriada, no algoritmo das formigas, para avaliar o risco médio do erro de classificação de modo a guiar o processo de seleção de variáveis em análise discriminante linear (LDA).
- Implementar o código do programa do algoritmo proposto ACO-LDA em linguagem MatLab ® 2011b.
- Demonstrar a eficiência do ACO-LDA em três estudos de caso com diferentes técnicas:
 1. Classificação de quatro tipos de óleos vegetais comestíveis (soja, milho, canola e girassol) utilizando a espectrometria de absorção molecular UV-Visível;
 2. Classificação de amostras de chás com respeito as suas variedades (preto e verde) e origens geográficas (argentino, brasileiro e srilanquense) utilizando a espectrometria de infravermelho próximo (NIR);
 3. Classificação de sementes de feijões segundo o tipo de cultivar (perola, pontal e transgênico) utilizando um espectrômetro portátil DLP NIRscan Nano EVM.
- Realizar um estudo comparativo através de diferentes parâmetros de desempenho (sensibilidade (SEM), especificidade (ESP) e acurácia(TCC)) entre o ACO-LDA e outros algoritmos populares e/ou bem estabelecidos na literatura (AG-LDA, PLS-DA e SPA-LDA).

CAPÍTULO II

FUNDAMENTAÇÃO

TEÓRICA



2. FUNDAMENTAÇÃO TEÓRICA

2.1 Dados em química analítica

Na química analítica, a quantidade de dados, objetos e variáveis muito grande, torna-se absolutamente indispensável a disposição ordenada dos dados em forma de matrizes para tornar mais fácil a sua manipulação. Dados multivariados são, em geral, organizados em matrizes através de vetores em linha ou coluna.

Uma organização adequada e avaliação dos dados são essenciais para que uma análise seja feita de forma correta. Para isso, é necessário ter o cuidado em investigar a forma como os dados foram gerados, os métodos utilizados, a confiabilidade destes dados e o mais importante, a representabilidade do grupo de amostras disponíveis que deseja analisar. De nada adiantaria um conjunto amplo de dados com informações interessantes se o grupo de amostras não fosse representativo o suficiente para fornecer as informações adequadas ao tratamento que se propõe fazer.^{31,32} De acordo com a dimensionalidade de informação gerada por amostras^{33,34}, os dados são categorizados^{35,36} como: ordem zero (univariada), primeira ordem (multivariada) e multivias (multidimensionais).

Dentre esses tipos de dados analíticos mencionados anteriormente, o foco do trabalho foi usar técnicas espectroscópicas que gerassem dados de primeira ordem. Então, apenas esses tipos de dados foram explorados ao longo do trabalho.

2.1.1 Métodos de Análise Multivariada

A aquisição de informações dos resultados de um experimento químico envolve a análise de um grande número de variáveis (dados de natureza multivariada). Os dados multivariados são caracterizados por apresentarem muitas informações ou medidas por cada amostra analisada. Muitas vezes, um pequeno número destas variáveis contém as

informações químicas mais relevantes, enquanto a maioria das variáveis adiciona pouco ou nada à interpretação dos resultados em termos químicos.^{31,32,37}

Devido a grande quantidade de variáveis geradas, torna-se necessária a aplicação de ferramentas que possibilitem a interpretação dos resultados. Para essa finalidade pode-se recorrer às técnicas disponibilizadas na Quimiometria.

Como citado por Kowalski³¹:

"as ferramentas quimiométricas são veículos que podem auxiliar os químicos a se moverem mais eficientemente na direção do maior conhecimento". Isto nos leva a uma definição formal de quimiometria: "... uma disciplina química que emprega métodos matemáticos e estatísticos para planejar ou selecionar experimentos de forma otimizada e para fornecer o máximo de informação química com a análise dos dados obtidos"³¹.

Portanto, para o tratamento de dados químicos é empregado de estatística multivariada, que é a quimiometria, com intuito de extrair deles a máxima informação química possível. E quando empregados em dados multivariados com o intuito de buscar tendências (padrões) e/o agrupamentos que buscam encontrar similaridade e diferença entre um dado conjunto de amostra e variáveis que foram submetidos a algum tipo de medida, essa atividade é realizada utilizando técnicas de classificação de reconhecimento de padrões supervisionadas ou não supervisionadas com a finalidade qualitativa ³⁸. A descrição desses métodos é apresentada na Seções **2.2.3.1 e 2.2.3.2**.

Antes de aplicar os métodos de reconhecimento de padrões, torna-se importante avaliar o uso de técnicas de pré-processamento dos dados. O uso inapropriado ou a ausência dessa etapa inicial poderá prejudicar o desempenho dos métodos adotados.

2.2.2 Pré-Processamento dos dados

As técnicas espectrométricas poderá apresentar sinal analítico, em alguns casos, intensidade com magnitudes diferentes e variações sistemáticas da linha de base e/ou ruído instrumental ³⁹. Então, antes mesmo de se aplicar os métodos de reconhecimento de padrões, o emprego de técnicas de pré-processamento dos dados deve ser previamente analisado.

A etapa de pré-processamento dos dados é fundamental para o sucesso da análise multivariada. Os principais objetivos da aplicação das técnicas de pré-processamento são eliminar informações não relevantes do ponto de vista químico e tornar a matriz de dados melhor condicionada para a análise, possibilitando a subsequente análise exploratória do conjunto de dados com eficiência ⁴⁰.

Os pré-processamentos dos dados são frequentemente empregados no domínio das amostras e/ou variáveis. A centralização dos dados na média, escalonamento e autoescalonamento são exemplos de pré-processamentos bastante utilizados nas variáveis⁴¹. Alguns pré-processamentos aplicados às amostras se destacam, tais como: variação normal padrão (standart normal variate –SNV), correção de espalhamento multiplicativo ou correção de sinal multiplicativo (multiplicative scatter correction - MSC), suavização, derivação, entre outros ^{31,42, 43}.

SNV, um auto-escalonamento no domínio das amostras é realizado, corrigindo os efeitos de espalhamento indesejado da luz e do tamanho das partículas ⁴⁴. A **Equação 1** mostra a transformação utilizada pela SNV ⁴⁵.

$$SNV_A = \frac{\sum_{i=1}^{\rho} (X_i - \bar{X})}{\sqrt{\frac{\sum_{i=1}^{\rho} (X_i - \bar{X})^2}{\rho - 1}}} \quad (1)$$

Onde, SNV_A são as variáveis normais padrão de ρ comprimentos de onda para uma amostra A; X é o valor do sinal analítico em i comprimento

de onda da amostra A e $\bar{X}$ é a média dos valores de ρ comprimentos de onda da amostra A, calculado conforme a **Equação 2**.

$$\bar{X} = \frac{\sum_{i=1}^{\rho} X_i}{\rho} \quad (2)$$

Já o MSC é um método de transformação que busca minimizar a influência de efeitos de espalhamento aditivos e multiplicativos nos espectros, tais como: o tamanho de partícula, a rugosidade, opacidade, sensibilidade do detector e do amplificador detector, variações na temperatura e pressão, os quais não trazem informações químicas sobre as amostras e introduz variações espectrais não relacionadas às informações químicas, como o deslocamento da linha de base ⁴⁶. Este método tenta remover o efeito do espalhamento pela linearização de cada espectro(X_i) por um espectro ideal. Para efeito de cálculo, considera-se que o espectro ideal é o espectro médio do conjunto de dados para o qual se deseja realizar a correção da linha base ⁴⁷. Em seguida, utiliza-se uma regressão linear para calcular o coeficiente angular e linear do gráfico entre o espectro ideal e o espectro que vai ser corrigido conforme a **Equação 3** ⁴⁷. Os coeficientes a_i e b_i são estimados por quadrados mínimos fazendo-se a regressão de cada espectro i no espectro média.

$$X_i = a_i \bar{X} + b_i \quad (3)$$

O espectro corrigido é calculado subtraindo cada ponto do espectro pelo valor do coeficiente linear e dividindo este valor pelo coeficiente angular conforme mostrado na **Equação 4** ^{47,48}

$$MSC = \frac{\sum (X_i - a_i)}{b} \quad (4)$$

Para resolver os problemas com a variação da linha de base, os espectros podem ser corrigidos através do método da 1ª derivada ou 2ª derivada que possibilita, uma melhor visualização dos picos existentes nos sinais originais ⁴⁴. A filtragem de ruído aleatório e aumento da relação sinal/ruído pode ser alcançada através da aplicação de procedimentos de suavização dos dados. Destaca-se se o método de Savitzky-Golay ⁴⁹ que ajusta um polinômio de baixa ordem aos pontos de uma janela pelos mínimos quadrados. Esses procedimentos permitem corrigir as informações contidas ao longo dos diferentes comprimentos de onda ⁴⁴.

2.2.3 Reconhecimento de Padrões

As técnicas de reconhecimento de padrões são usadas para identificar as semelhanças e diferenças em diferentes tipos e amostras que foram submetidas a algum tipo de análise, seja por técnicas instrumentais (espectrométricas, eletroanalíticas, cromatográficas, entre outras), seja pela determinação de alguns parâmetros de interesse da amostra (pH, densidade, concentração de algumas espécies, etc) ^{50,51}. Destacar uma característica ou encontrar tendências/agrupamentos em conjunto dados torna-se uma tarefa cada vez mais difícil à medida que o número de amostras e variáveis aumenta. Para essa tarefa, técnicas de Reconhecimento de Padrão (RP) têm sido cada vez mais relatadas na literatura ^{32,52}. Estas técnicas se fundamentam nas seguintes suposições para encontrar os critérios de classificação: amostras do mesmo tipo são semelhantes, existem diferenças entre diferentes tipos de amostras e as semelhanças e diferenças são refletidas nas medidas utilizadas para caracterizar as amostras.

As técnicas de Reconhecimento de Padrão RP são divididas em supervisionada e não supervisionada (análise exploratória).

2.2.3.1 Técnicas de Reconhecimento de Padrões não Supervisionadas

As técnicas de reconhecimento de padrões não supervisionadas buscam encontrar em um conjunto de dados, tendências e/ou agrupamentos sem a necessidade do conhecimento prévio das classes. Essas técnicas utilizam apenas medidas de algumas propriedades com finalidade de demonstrar agrupamentos naturais.

Técnicas de reconhecimento de padrão não supervisionadas são amplamente encontradas na literatura, onde podemos destacar a análise de agrupamentos hierárquicos (HCA, do inglês *Hierarchical Cluster Analysis*) e análise de componentes principais (PCA, do inglês *Principal Component Analysis*) ⁴¹. Esses métodos são complementares, com muita aceitação por parte dos pesquisadores nas análises de dados químicos. Neste trabalho, apenas a PCA foi avaliada.

2.2.3.1.1 Análise por Componentes Principais-PCA

A PCA é uma das técnicas de RP não supervisionadas que consiste essencialmente em reescrever as coordenadas das amostras em outro sistema de eixo mais conveniente para a análise dos dados. Ou seja, combinações lineares das variáveis originais são empregadas com o propósito de gerar um novo sistema de eixos ortogonais entre si e não correlacionados, denominados de componentes principais (PC, do inglês *Principal Components*) ⁴¹. A informação contida em uma delas não está presente em outra. Essas PCs são obtidas com respeito a quantidade de informações dos dados originais que cada uma dessas novas variáveis é capaz de descrever. A primeira delas, PC1, é definida pela direção que descreve a máxima variância dos dados originais. A segunda, PC2, tem a direção de máxima variância dos dados no subespaço ortogonal à PC1, e as componentes subsequentes são ortogonais às anteriores e orientadas de maneira que descrevam, uma a uma, a máxima variância restante ⁴⁷. Portanto a ordem é decrescente de máxima variância, ou seja, a PC1

detém mais informação estatística que a PC2, que por sua vez tem mais informação estatística que a PC3 e assim por diante ^{53,54}. Assim, este novo conjunto de eixos de coordenadas no qual projetaram-se as amostras, escores, é muito mais informativo e, pelo fato de serem ordenados pela sua importância, é possível visualizar estas mesmas amostras num gráfico de baixa dimensionalidade ⁵³. Nesse conjunto, apenas as variáveis realmente representativas estão presentes, o que funciona como uma espécie de filtro para as informações redundantes ou de caráter ruidoso. E os coeficientes da combinação linear, os pesos, correspondem a uma medida da contribuição de cada variável original ^{53,55}. Matematicamente, na PCA, a matriz X é decomposta em um produto de duas matrizes, denominadas escores (T) e pesos (P), mais uma matriz de erros (E), como mostrado na **Equação 5**:

$$X = TP + E \quad (5)$$

A **Figura 2.1** é uma representação gráfica de um conjunto de amostras no espaço bidimensional definido por duas variáveis e as respectivas componentes principais, PC1 e PC2. A linha tracejada de cor preta mostra as coordenadas de uma amostra em relação aos dois eixos das variáveis originais, e a linha tracejada verde mostra as coordenadas dessa amostra em relação aos primeiros eixos das componentes principais ⁵⁵.

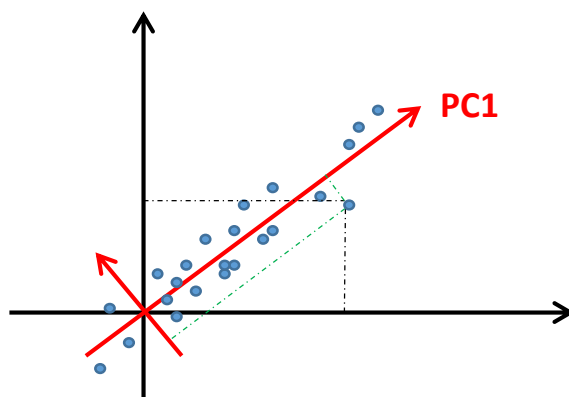


Figura 2.1. PC1 e PC2 obtidas a partir da combinação linear das variáveis originais. As linhas tracejadas na cor verde correspondem aos escores e as linhas tracejadas na cor pretas representam os pesos ⁵⁵.

A PCA pode ser utilizada para ^{53,55}:

- Observar agrupamentos no conjunto de dados;
- Verificar a contribuição das variáveis originais no novo conjunto de dados obtidos;
- Identificar amostras anômalas (*outliers*).

2.2.3.2 Técnicas de Reconhecimento de Padrões Supervisionadas

Diferentemente das técnicas de reconhecimentos de padrões não supervisionados, as supervisionadas utilizam uma informação adicional sobre membros (objetos) das classes. Assim a resposta instrumental é relacionada com os índices de classes e modelos multivariados são obtidos e validados com intuito de classificar amostras de identidades desconhecidas⁵⁵⁻⁵⁷. A **Figura 2.2** ilustra um esquema de como organizar os dados para implementação de uma análise multivariada (classificação).

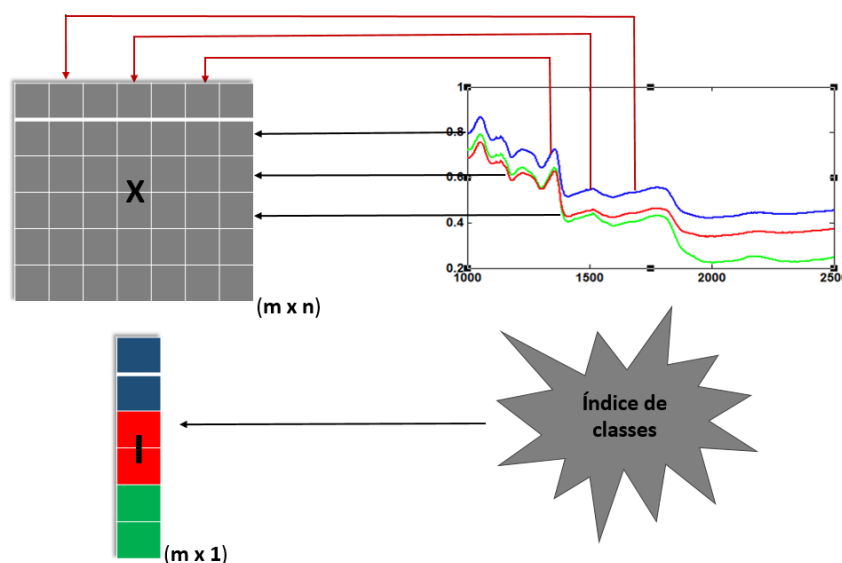


Figura 2.2. Organização dos dados para análise multivariada (classificação)⁴⁰.

Na literatura são encontrados diversos métodos de reconhecimento de padrões supervisionado em problemas analíticos de classificação, destacando a Modelagem Independente e Flexível Por Analogia de Classes (SIMCA, do inglês *Soft Independent Modeling of Class Analogy*), a Análise

Discriminante Linear (LDA, do inglês *Linear Discriminant Analysis*) e os mínimos quadrados parciais para análise discriminante PLS-DA. Nesse trabalho serão utilizados o PLS-DA e a LDA.

2.2.3.2.1 Análise Discriminante pelos Mínimos Quadrados Parciais-PLS-DA

O método dos mínimos quadrados parciais (PLS, do inglês Partial Least Squares), amplamente conhecido e utilizado em diversas ciências, não foi concebido como um método de classificação, mas sim regressão. E tem sido usado como ferramenta de regressão na Quimiometria há mais de 30 anos ⁵⁷. PLS é um método de projeção multivariada para modelar uma relação entre variáveis independentes (X) e variáveis dependentes (Y). E seu princípio é encontrar as componentes na matriz de entrada (X) que descreva tanto quanto possível as variações significativas nas variáveis de entrada e, ao mesmo tempo, tem correlação máxima com o valor alvo em Y. E assim, encontram simultaneamente as variáveis latentes em X que irão prever as variáveis latentes em Y. ⁵⁸ Em análise multivariada, a técnica de classificação PLS-DA, combina as propriedades da Regressão por Mínimos Quadrados Parciais (PLS) com a habilidade de classificação da Análise Discriminante (DA). No PLS-DA são encontradas variáveis latentes (VL) ou componentes principais (PCs), que resultam da combinação linear das variáveis, que tem o objetivo de obter uma melhor discriminação das classes de amostras baseado na covariância máxima entre a matriz de dados instrumentais e uma matriz Y ⁵⁸⁻⁶⁰. O número ótimo de variáveis latentes pode ser estimado por meio de validação cruzada ou conjunto de teste externo. ^{61,58-60,62}

Em PLS-DA a matriz Y é composta por valores binários, comumente são utilizados 0 ou 1⁶³. Por exemplo, se temos L classes envolvidas ($l=1,2,3,\dots,L$), a matriz Y tem dimensões ($I \times L$). Para a primeira classe, correspondente a $l=1$, o número (1) é atribuído as amostras dessa classe, e o zero (0) é atribuído a todas as outras amostras que não pertence a essa

classe. Já na próxima coluna da matriz Y, quando $l=2$, são as amostras dessa classe que são designadas por um, e todas as outras por zero, e assim por diante até que cada uma das classes tenha sido representada pelo número 1⁴⁷. A classificação continua obtendo-se valores entre 0 e 1 para cada objeto a ser discriminado, sendo necessário estabelecer um valor que define a fronteira entre as classes. A fronteira, denominada threshold, pode ser calculada pelo teorema de Bayesian ou através dos limites de confiança para cada objeto, sendo comumente adotado o valor de 0.5⁶⁴. A **Figura 2.3** ilustra a definição da matriz Y para um conjunto de dados de três variáveis.

Matriz X	Matriz Y		
	Y ₁	Y ₂	Y ₃
CLASSE 1 ● ● ●	1	0	0
	1	0	0
	1	0	0
CLASSE 2 ★ ★ ★	0	1	0
	0	1	0
	0	1	0
CLASSE 3 ■ ■ ■	0	0	1
	0	0	1
	0	0	1

Figura 2.3. Representação do modelo preditivo PLS-DA.

Uma sequência de operações matemáticas decompõe a matriz X, formada pelas variáveis de resposta instrumental, em variáveis latentes, a fim de reduzir a dimensionalidade do conjunto de dados.

O PLS-DA é um método de regressão combinado com uma técnica de classificação. Diferente das metodologias de reconhecimento de padrões não supervisionados (PCA), PLS-DA maximiza a separação entre classes pré-definidas, em vez de explicar as variações dentro de um conjunto de

de dados ⁶⁵. Na PLS-DA, as PCs do PCA que eram ortogonais, agora são conhecidas como variáveis latentes e não são ortogonais.

Para problemas de classificação envolvendo mais de duas classes, o algoritmo PLS2 é utilizado e o número de variáveis de resposta é correspondente ao número de classes envolvidas ⁶⁶.

2.2.3.2.2 Análise Discriminante Linear-LDA

A análise de discriminantes lineares (LDA), também conhecidos como discriminantes lineares de Fisher ⁶⁷, é uma técnica baseada no cálculo de hiperplanos, ou funções discriminantes (FD) lineares. Essas funções maximizam a separabilidade entre as classes e minimizam a variância entre os objetos (ou amostras) de uma mesma classe. Para isso, as FD lineares são obtidas por meio de combinações lineares das variáveis originais que melhor discriminem as classes ou grupos de objetos ⁵¹. Em outras palavras, a abordagem de Fisher transforma os dados químicos de natureza multivariada em uma variável latente de forma a separar os grupos transformados o máximo possível ⁶⁸. Quando um estudo de classificação apresentar c classes de objetos (ou amostras), $c - 1$ funções discriminantes poderão ser determinadas se o número de variáveis for maior do que " c " ⁵¹.

Uma das maneiras mais simples de se obter uma função linear discriminante é ilustrada na **Figura 2.4**, onde os centros de gravidade das classes formadas por círculos e quadrados são inicialmente determinados. Em seguida traça-se um plano de decisão equidistante e perpendicular à linha que une os centros de gravidade dos grupos, formando as regiões R_1 e R_2 . O ideal é que as amostras de cada classe sejam semelhantes entre si e estejam distantes das amostras da outra classe. Se um ponto qualquer estiver presente na região R_1 , o mesmo é classificado como pertencente à classe 1. Caso contrário, obrigatoriamente pertencerá à classe 2 ⁶⁹.

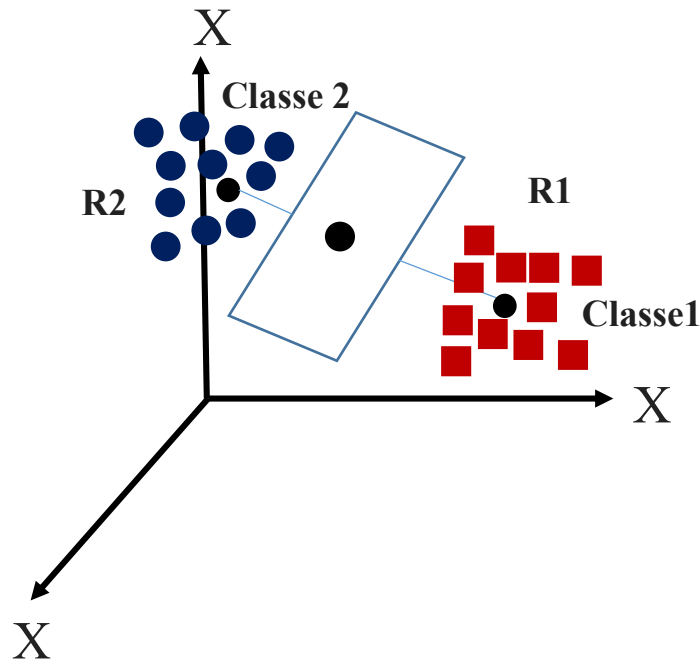


Figura 2.4. Representação esquemática da análise discriminante linear.

O critério para separação dos grupos baseia-se no cálculo da distância de Mahalanobis ^{42,55} sendo uma amostra descrita por $\mathbf{x} = [x_1, x_2 \dots x_m]^T$, em que $[x_1, x_2 \dots x_m]$ corresponde às variáveis referentes à técnica aplicada, como as medidas de absorvância em diferentes comprimentos de onda. O quadrado da distância de Mahalanobis $r^2(\mathbf{x}, \mu_j)$ entre $\mathbf{x}$ e o centro da j -ésima classe ($j=1,2,\dots, c$), será definido conforme **Equação 6**:

$$r^2(\mathbf{x}, \mu_j) = (\mathbf{x} - \mu_j)^T \cdot \sum_j^{-1} (\mathbf{x} - \mu_j) \quad (6)$$

A LDA tem sua funcionalidade restrita a conjuntos de dados de pequenas dimensões. Além disso, a capacidade de generalização de modelos LDA pode ser comprometida por problemas de multicolinearidade ⁵¹. A multicolinearidade pode ser explicada como uma inter-relação entre as variáveis independentes⁷⁰, assim, tais variáveis podem fornecer informações redundantes. Quando isso ocorre a habilidade de generalização do modelo LDA pode ser comprometida⁵¹. Dessa forma, o uso da LDA para

classificação, sobretudo com dados espectrométricos, necessita de procedimentos de redução de dimensionalidade e/ou seleção de variáveis. A utilização de algoritmos de seleção de variáveis tem sido utilizada com sucesso [34, 71](#).

2.3. Seleção de variáveis

A escolha da região espectral é muito importante para a eficiência de um modelo multivariado. Quando escolhida adequadamente, permite minimizar os erros de predição de forma a obter modelos mais robustos, de simples interpretação e a aumentar o desempenho de alguns métodos de reconhecimento de padrões e de regressão pela retirada de muitas variáveis irrelevantes ou com informações redundantes, além de minimizar a multicolinearidade [71,72](#). Essas características são comuns das técnicas mais utilizadas atualmente como as espectroscópicas que apresentam em controvérsia a estes problemas, vantagens como sensibilidade, rapidez e uso reduzido (ou nenhum uso) de reagentes contribuindo para a química verde. Ademais, a seleção de variáveis favorece a interpretação química dos espectros por operar no domínio dos dados originais, bem como possibilita o desenvolvimento, a partir de variáveis selecionadas, de instrumentos portáteis para serem usados em campo. Por isso as metodologias de seleção de variáveis foram desenvolvidas.

A seleção de variáveis é amplamente empregada e difundida em diversos campos da quimiometria, como por exemplo, classificação [73](#), calibração multivariada [74](#) e transferência de calibração [75,76](#) para os mais diversos métodos de regressão como MLR⁷⁷, PLS⁷⁸, dentre outros. As técnicas de seleção de variáveis são categorizadas de acordo com a forma que o subconjunto de variáveis selecionado assume: técnica seleção de variáveis individuais e intervalos [49,79](#), ou quando estão relacionadas com a presença de variáveis aleatória: podendo ser randômicas ou determinísticas [79,80](#).

A técnica de seleção de variáveis individuais seleciona um conjunto de variáveis que detém as melhores informações dentro o espectro

completo. Já por intervalos, consiste em encontrar uma região do espectro (faixa) que produza melhores resultados que o espectro completo. Para atingir este objetivo, o espectro é dividido em iguais intervalos, a quantidade de intervalos é definida pelo usuário.

Os métodos determinísticos não estão associados a nenhuma probabilidade a priori, apresentando como solução um único subconjunto de variáveis ⁴⁰. Tal método sempre leva à mesma resposta, após diferentes execuções, desde que aplicados sempre às mesmas condições iniciais. Ou seja, tem a reprodutibilidade do subconjunto de variáveis selecionado. Isso é a vantagem destes algoritmos em relação aos estocásticos. Porém, como nesses métodos a solução encontrada é extremamente dependente do ponto de partida fornecido, pode-se convergir para um ótimo local, não possuem bom desempenho em otimizar funções multimodais, isto é, funções que possuem vários ótimos locais ⁸⁵. Como exemplo, pode-se destacar o uso da seleção de variáveis na busca por um subconjunto que leve a um melhor modelo a partir de dados multivariados.

Um modo de amenizar o problema de ótimos locais é inserir no algoritmo um componente estocástico, tornando o algoritmo probabilístico. Os métodos randômicos, possuem alguma variável de entrada aleatória e, conseqüentemente, o subconjunto de variáveis selecionado é estocástico, ou seja, está associado a um certo grau de probabilidade ⁴⁰. Nestes métodos, várias escolhas são feitas com base em números aleatórios, sorteados no momento de execução do código. Como a cada execução do código os números sorteados serão diferentes, um método aleatório não executará a mesma sequência de operações em duas execuções sucessivas. Partindo de um mesmo ponto inicial, cada execução do código seguirá o seu próprio caminho, e possivelmente levará a resultados finais que geralmente convergem para as mesmas soluções ótimas com dada precisão ⁸¹.

Diversos métodos têm sido propostos para a finalidade de seleção de variáveis, entre os quais são descritos, a seguir, o determinístico (algoritmo

de projeções sucessivas - SPA) e o estocástico bioinspirado (algoritmo genético-GA).

2.3.1 Algoritmo das Projeções Sucessivas.

Araújo e colaboradores ⁸² desenvolveram um método de seleção de variáveis denominado de Algoritmo das Projeções Sucessivas (SPA), cujo objetivo é reduzir a colinearidade de um conjunto de dados de modelos MLR, melhorando assim, a capacidade preditiva desses modelos. E posteriormente, o SPA foi estendido para selecionar intervalos de variáveis em PLS ^{78, 83}, bem como para resolver problemas de multicolinearidade em análise discriminante linear, SPA-LDA ⁸⁴.

O princípio do algoritmo é baseado numa busca sistemática de variáveis informativas de modo que a operação seja iniciada partindo-se de um comprimento de onda ou vetor inicial escolhido, e em seguida, outros comprimentos de onda (vetores), não-colineares ao vetor escolhido, são incorporados um a um a cada iteração, em projeções sucessivas, até que um número específico de comprimentos de onda seja alcançado ⁸⁵.

No contexto da calibração multivariada em MLR o SPA envolve essencialmente três etapas. Na primeira etapa, operações de projeções são empregadas nas colunas da matriz de calibração, e um subconjunto de variáveis é formado levando em consideração o critério de minimização da multicolinearidade ⁸². Na segunda fase, é escolhido o subconjunto que obtiver melhor resultado em relação ao critério que avalia a habilidade de previsão de um modelo MLR, a Raiz Quadrada do Erro Médio Quadrático de Validação (RMSEV, do inglês *Root Mean Square Error of Validation*) ⁷⁴, de acordo com a **Equação 7**.

$$RMSEV = \sqrt{\frac{1}{K_v} \sum_{k=1}^{K_v} (y_v^k - \hat{y}_v^k)^2} \quad (7)$$

onde: K_v é número de amostras do conjunto de validação e y_v^k e $\hat{y}_v^k$ são os valores de referência e os valores previstos para o parâmetro de interesse nas amostras de validação.

Na terceira fase, o subconjunto escolhido é submetido a um procedimento de eliminação para determinar se alguma variável pode ser removida sem perda significativa da capacidade de predição ⁸².

Nos métodos de reconhecimento de padrões, a informação da modelagem está contida nos dados oriundos da resposta instrumental e no índice de classes para cada amostra. Nesse caso, o SPA possui uma versão aplicada a LDA ^{84, 86, 87}, a diferença principal estar na função de custo utilizada ao invés RMSEV para guiar a classificação, utiliza o risco médio Gda LDA, conforme a **Equação 8** ⁵¹.

$$G = \frac{1}{K_v - L - C} \sum_{k=1}^{K_{train}} g_K \quad (8)$$

Onde g_K é o risco de uma classificação incorreta da amostra de validação e é definido na **Equação 9**. E o seu valor deve assumir valores muito pequenos quanto à amostra estiver perto do centro da sua verdadeira classe e distante dos centros das demais classes ⁸⁰.

$$g_K = \frac{r^2(X_K, \mu_{IK})}{\min_{j \neq IK} r^2(X_K, \mu_{IJ})} \quad (9)$$

Na **Equação 9**, o numerador $r^2(\mathbf{X}_k, \mu_{IK})$ é o quadrado da distância de Mahalanobis entre o objeto $\mathbf{X}_k$ (com índice de classe I_k) e a média de sua classe (μ_{IK}). O denominador corresponde ao quadrado da distância de Mahalanobis entre o objeto $\mathbf{X}_k$ e o centro da classe errada mais próxima ^{51,84}. Com isso, menor risco será obtido quanto mais distante o objeto estiver do centro da classe incorreta e mais próximo estiver do centro de sua classe verdadeira ⁵¹.

Uma alternativa para problemas de classificação foi proposta por Soares et al. (2014) ⁸⁸ envolvem um numero reduzido de amostras. Nessecaso, o conjunto de validação foi dispensado e o risco de uma classe incorreta é calculada usando os mesmos objetos de treinamento que foram empregados para calcular as medias de cada classe e uma matriz de covariância. A função e custo G_{train} correspondente a essa adaptação é apresentada na **Equação 10**.

$$G_{train} = \frac{1}{K_{train}} \sum_{k=1}^{K_v} g_{train,k} \quad (10)$$

Na **Equação 10**, K_{train} é o número de amostras de treinamento, L é o numero de variáveis no subconjunto que está sendo avaliado e C representa o número de classes investigadas. Essa expressão evita possíveis problemas de sobreajuste relacionados ao uso repetido do conjunto de treinamento ⁸⁸.

2.3.2 Algoritmo genético

O Algoritmo genético foi proposto, na década de 60, pelo cientista John H. Holland da Universidade de Michigan, onde a ideia matemática principal simulava o mecanismo da “Teoria da Evolução das Espécies” de Charles R. Darwin ⁸⁹. Ou seja, buscam reproduzir o mecanismo biológico da evolução das espécies.

O funcionamento do algoritmo genético baseia-se no processo evolutivo dos seres vivos, seguindo o princípio básico de que as gerações futuras (descendentes) serão mais “evoluídas” do que os seus antecedentes ¹⁶. Onde gerações melhores continuariam existindo, enquanto que gerações mais “frágeis” tenderiam a sucumbir. Baseando-se neste princípio de geração de descendentes mais aptos (evoluídos), o algoritmo genético básico interpreta as informações dos sistemas como cromossomos

genéticos e possui os seguintes passos: 1.codificação das variáveis, 2.geração da população inicial, 3.avaliação das aptidões, 4.cruzamento e 5.mutação.

O primeiro passo consiste em codificar as variáveis de modo que estas se assimilem a cromossomos biológicos. Neste contexto, o procedimento mais comum é usar codificação binária em que cada variável representa um gene e pode receber valor 0 ou 1, o que significa respectivamente não incluída e incluída no modelo ⁴⁰. Na **Figura 2.5** é apresentado um esquema da codificação binária.

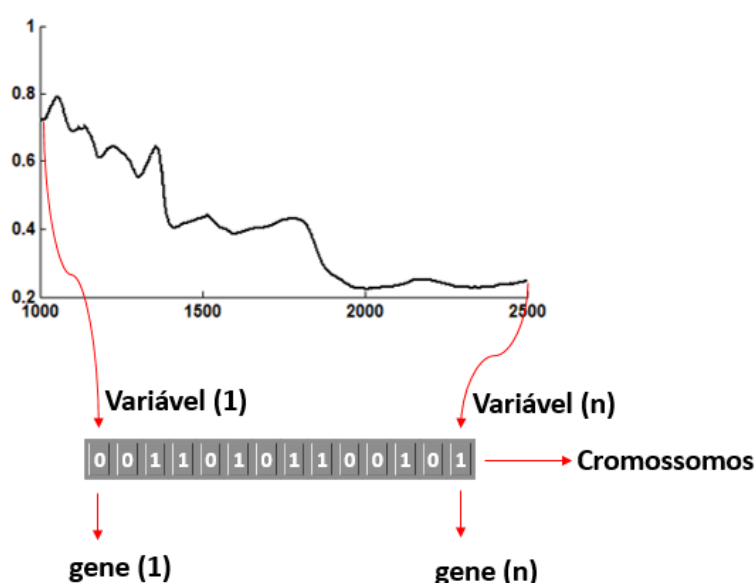


Figura2.5. Codificação binária usada na seleção de variáveis do GA.

Como observado na **Figura 2.5** cada coluna em cinza representa um indivíduo da população inicial, ou seja, um gene. E um conjunto de genes resulta no cromossomo. Assim, no caso de problemas de seleção de variáveis em química analítica, cada gene é indicado por uma variável e o conjunto das variáveis resulta no cromossomo, onde cada comprimento de onda (variável) de um espectro é representado por um gene que pode ou não ser selecionado ^{90,91}. Ou seja, o tamanho do cromossomo corresponde a todos os comprimentos de onda do espectro selecionado.

O segundo passo consiste na geração da população inicial de cromossomos (indivíduos). E para que isso ocorra, o algoritmo é executado inicialmente de forma randômica para a formação de uma população de cromossomos que representem possíveis soluções do problema, como mostrado na **Figura 2.6**. Essa geração é completamente aleatória e cada indivíduo gerado é avaliado na próxima fase de execução do algoritmo.

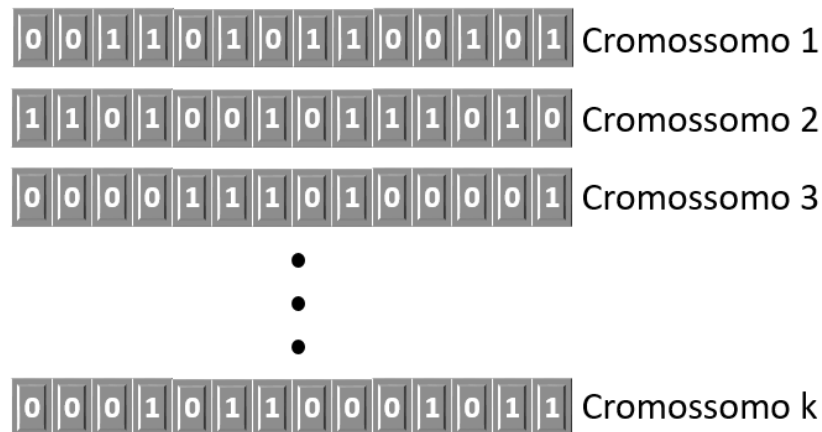


Figura 2.6 Representação dos cromossomos gerados de maneira aleatória.

O terceiro passo consiste em avaliar a aptidão de cada indivíduo, o que na teoria da evolução corresponderia à capacidade de sobreviver. Matematicamente, representa a capacidade de gerar melhores respostas, sendo que, no contexto de seleção de variáveis, quando maior a aptidão menor o erro de predição obtido ⁴⁰. E este será o responsável direto pela presença ou ausência de uma determinada característica dos indivíduos. A aptidão é calculada como o inverso do risco médio de classificação, definido como no SPA-LDA, nas modelagens de análises multivariadas. Os que apresentarem boa aptidão são selecionados para próxima fase do algoritmo, a reprodução. O critério de escolha é conhecido como método da roleta ⁹², sendo que a probabilidade de cada indivíduo gerar descendentes é proporcional a sua aptidão em sobreviver. Neste método, cada indivíduo da população é representado na roleta proporcionalmente ao seu índice de aptidão. Assim, aos indivíduos com alta aptidão é dada uma porção maior

da roleta, enquanto aos de aptidão mais baixa é dada uma porção relativamente menor da roleta. Finalmente, a roleta é girada um determinado número de vezes, dependendo do tamanho da população, e são escolhidos, como indivíduos que participarão da próxima geração, aqueles sorteados na roleta.

No quarto passo do algoritmo é realizado o cruzamento entre material genético de diferentes indivíduos. A nova população é formada a partir do cruzamento aleatório de pares de cromossomos e da posição onde ocorrerá a ruptura seguida pela troca de material genético. O número de rupturas geralmente varia entre 1 e 2 ⁴⁰. Isto se deve ao fato de que quanto maior for o número de rupturas menor será a semelhança entre os pais e os filhos, portanto dificultando a convergência. Assim, durante o processo de permuta do material genético, a tendência natural é que, após certo número de gerações, as características dominantes comecem a prevalecer na população, levando assim à convergência do algoritmo ³⁷. A **Figura 2.7** mostra de forma esquemática o passo do cruzamento entre dois indivíduos

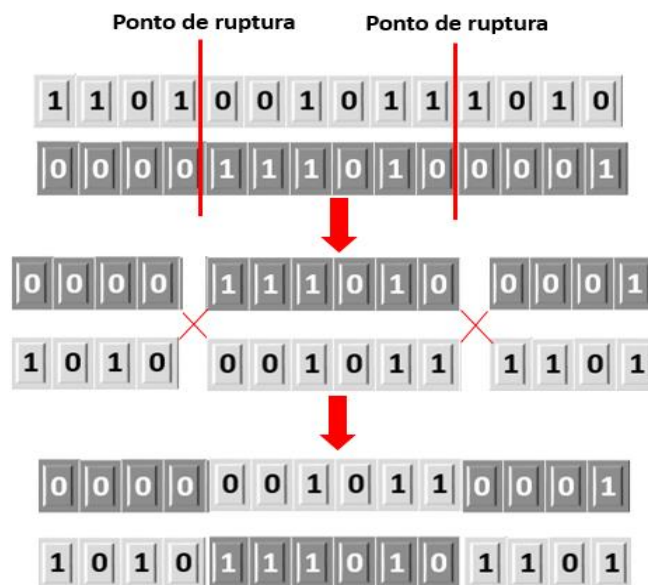


Figura 2.7. Esquema de processo de cruzamento no algoritmo genético.

No quinto passo, em conjunto com a reprodução, ocorre também o processo de mutação, que acontece devido a uma perturbação das informações contidas dentro dos genes durante o processo de reprodução.

A mutação promove alterações no código genético de uma pequena parcela da população, onde muitas vezes pode atuar de forma benigna ⁴⁰. Se a alteração gerar indivíduos mais adaptados, estas novas características serão transmitidas para os demais indivíduos ao longo das próximas gerações ⁴⁰. Este mecanismo, evita que o algoritmo fique preso a mínimos locais na otimização. Em termos práticos, a mutação consiste em substituir 0 ou 1 e vice e versa, como ilustrado na **Figura 2.8**.



Figura 2.8. Esquema de processo de mutação no algoritmo genético.

Ao final das n gerações escolhidas pelo usuário, o subconjunto de variáveis que produzir o melhor resultado com base no valor da função objetivo é a solução apresentada pelo GA.

Estes são os principais parâmetros do algoritmo genético básico. Para sua execução geralmente determina-se uma probabilidade de mutação baixa (abaixo de 0,5%), pois valores altos impediriam a passagem de genes dominantes ao longo das gerações tornando o algoritmo apenas aleatório o que dificultaria a convergência⁹³, e determina-se uma probabilidade de cruzamento geralmente de 0.6⁹⁴ que corresponde a 60%.

O GA foi utilizado na área de química, pela primeira vez, no trabalho desenvolvido por por Lucasius e Kateman⁹¹, onde váriavies foram selecionadas na região do ultravioleta para a determinação de nucleotídeos. Atualmente diversos trabalhos reportam a utilização do GA em diversas matrizes químicas devido a vantagens como⁴⁰:

- Não requer informação *a priori* sobre o gradiente de resposta;
- A existência de descontinuidade não afeta o desempenho do algoritmo;

- Existem operadores que evitam que o algoritmo fique preso a mínimos locais;

No presente trabalho um algoritmo estocástico, inspirado no comportamento da colônia de formigas, é proposto para seleção de variáveis em modelagem LDA para fins de classificação, cuja descrição será apresentada no capítulo III. Antes, descreveremos, a seguir, sobre colônias de formigas, otimização por colônia de formigas e sua aplicação na química.

2.3.3 Colônias de Formigas.

2.3.3.1 Colônias de Formigas reais

As formigas reais são capazes de encontrar o caminho mais curto para uma fonte de alimento do formigueiro sem a utilização de dados visuais⁹⁵. Ao saírem do formigueiro em busca de comida, as formigas depositam no chão uma substância, feromônio, que serve de comunicação entre elas para a formação de trilhas do ninho até a fonte de alimento⁹⁶. Estas trilhas de feromônios podem ser observadas por outras formigas e motivá-las em seguir determinado caminho, isto é, um movimento aleatório das formigas segue com maior probabilidade uma trilha de feromônio⁹⁶. Esta é uma maneira de como as trilhas são reforçadas e mais e mais formigas tendem a seguir aquela trilha mais rapidamente minimizando a distância conforme mostrada na **Figura 2.9**. Diante disso, o caminho mais curto entre o formigueiro-alimento é escolhido.

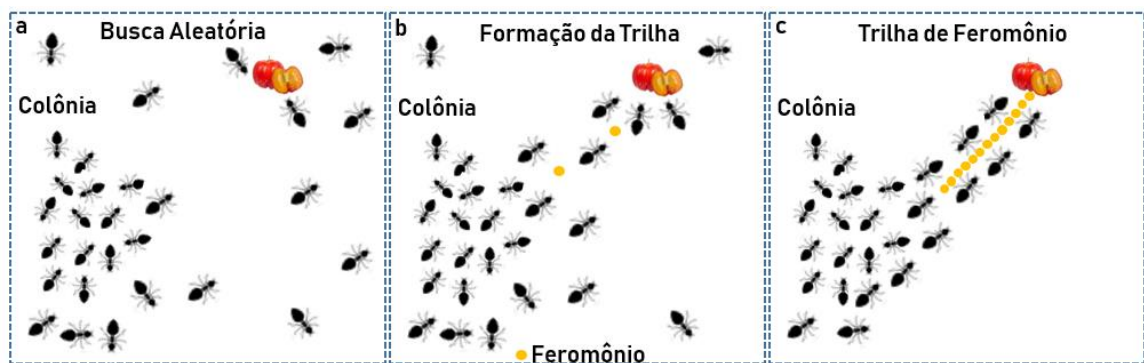


Figura 2.9. Formigas na busca do melhor caminho até o alimento. a) as formigas forrageiam aleatoriamente. b) formação da trilha de feromônio. c) trilha formada.

O estudo do comportamento forrageiro das formigas foi realizado por Deneubourg et al.⁹⁷, através de um experimento chamado de ponte binária conforme ilustrado na **Figura 2.10**. No início, as formigas são deixadas livres para escolher o caminho entre o formigueiro e a fonte de alimento. Assim, as formigas não têm uma preferência, movimentando-se com probabilidades iguais para qualquer um dos caminhos. O resultado observado foi que, apesar de inicialmente as formigas se movimentarem de forma aleatória, todas as formigas escolheram o mesmo caminho. Desse modo, a concentração de feromônio nesse caminho será mais forte, induzindo outras formigas a percorrerem o mesmo caminho, que por sua vez estimula mais formigas para escolher novamente aquele caminho, e assim por diante, até que as formigas tenham convergido para um único caminho ⁹⁸⁻¹⁰¹.

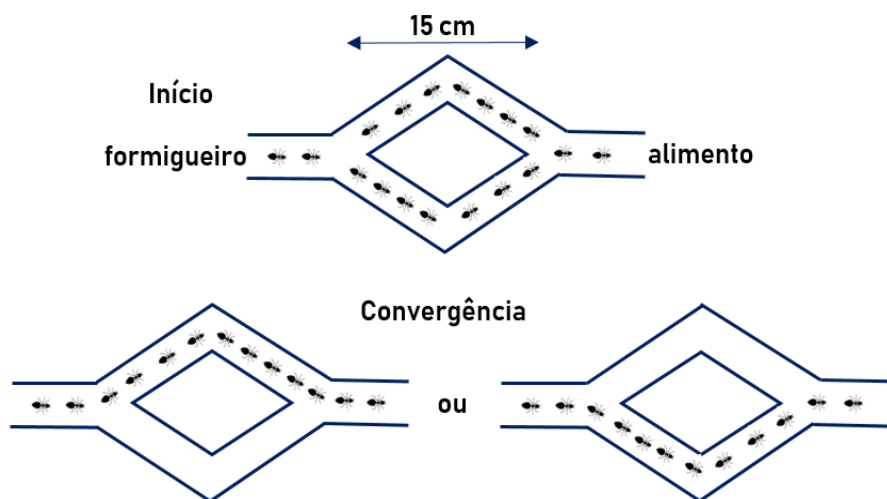


Figura 2.10. Formigas na busca do melhor caminho até o alimento. a) as formigas são deixadas livres para escolher o caminho. b) as formigas convergem para um dos caminhos probabilidade iguais ⁹⁸⁻¹⁰¹

Outra observação foi realizada usando pontes de tamanhos diferentes de modo que o maior caminho tem duas vezes o tamanho do caminho menor (**Figura 2.11**). Assim como no primeiro experimento, as formigas saem em busca de alimento e ao chegarem ao ponto de decisão elas escolhem aleatoriamente, pois inicialmente os ramos parecem idênticos às

formigas. Com base no experimento anterior, seria de se esperar que as formigas escolhessem um ou outro caminho, onde a escolha de um caminho para o outro dependeria de oscilações estocásticas. No entanto, não é isso que acontece e as formigas escolhem o caminho mais curto. Isso é justificado pelo fato de que as formigas que por um acaso escolheram o caminho mais curto retornam primeiro da fonte de alimento para o ninho. Logo, o caminho mais curto apresentará uma maior concentração de feromônio, estimulando mais formigas a seguirem esse caminho. Dessa forma, o feromônio começa a se acumular mais rapidamente no ramo mais curto, atraindo as formigas **98-101**.

Ao decorrer do tempo, o feromônio sofre o processo de evaporação, pois a substância é volátil, e a concentração em caminhos menos visitados vai diminuindo, reduzindo também, a influência desses caminhos nas decisões das formigas.

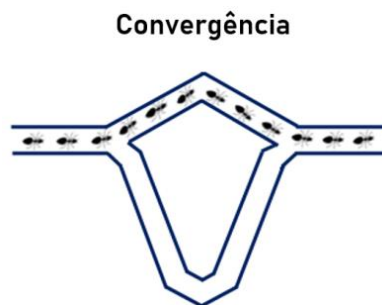


Figura 2.11. Comportamento das formigas que convergem para a ponte mais curta **98-101**.

Portando, os estudos sobre o comportamento das formigas reais levaram muitos pesquisadores a desenvolverem outra classe de algoritmos, usando formigas artificiais, chamada ACO (*Ant Colony Optimization* - Otimização por Colônia de Formigas) para resolver problemas combinatórios complexos e as informações geradas por instrumentos analíticos que portam de uma complexidade resultante da multiplicidade das variáveis, através do método de seleção de variáveis.

2.3.3.2 Colônias de Formigas artificiais

Formigas artificiais são caracterizadas como agentes que imitam o comportamento das formigas reais. E foram criadas para solucionar problemas de otimização ACO. Computando via feromônio à capacidade de se comunicar, marcar e rastrear as trilhas como formigas artificiais, o algoritmo de colônia de formigas foi criado por Dorigo.^{26,98,99,102}

O nível de feromônio artificial permite a formiga intensificar a busca por um determinado caminho, ou trilha por um circuito menor e gastando, portanto, menos tempo. Isto significa que, quanto maior a quantidade de feromônio na trilha, melhor é o caminho¹⁰⁰. A sensibilidade em relação à substância química permite uma comunicação interativa entre as formigas reais. A transposição matemática deste nível de controle ou sensibilidade de feromônio leva à construção da solução heurística de colônia de formigas artificiais para resolver problemas difíceis de otimização¹⁰¹. A análise probabilística do comportamento das formigas artificiais partiu do modelo estocástico do comportamento das formigas reais, conforme descrito por ⁹⁸⁻¹⁰¹.

As formigas “artificiais” possuem muitas semelhanças e algumas diferenças com relação às formigas “reais”, encontradas na natureza. No ACO são usadas formigas artificiais e que diferem das formigas reais em¹⁰³:

- As formigas artificiais têm memória;
- Não são completamente cegas;
- Elas vivem em um ambiente onde o tempo é discreto, enquanto nas formigas reais os movimentos são contínuos.

Por outro lado, diversas características são idênticas aos formigueiros reais¹⁰³:

- Formigas artificiais têm uma maior preferência probabilística por trilhas com maior quantidade de feromônio;
- Caminhos mais curtos tendem a terem maiores taxa de crescimento em termos de depósito de feromônio;
- As formigas artificiais usam um tipo de comunicação indireta, baseada na quantidade de feromônio depositada em uma trilha.

- Comportamento estocástico: a forma probabilística é característica das duas realidades.

2.3.4 Algoritmo de otimização por colônia de formigas

O algoritmo de otimização por colônia de formigas são modelos matemáticos baseados no comportamento das formigas, é um termo que representa uma classe de meta-heurísticas bio-inspiradas baseadas no comportamento das formigas. Desde a proposta do Ant System, o primeiro algoritmo ACO²⁶, muitos resultados significativos de pesquisa foram obtidos. Essas contribuições se concentraram no desenvolvimento de variantes algorítmicas de alto desempenho, no desenvolvimento de uma estrutura algorítmica genérica para o algoritmo ACO, em aplicações bem-sucedidas dos algoritmos ACO para uma ampla gama de problemas computacionalmente difíceis, e a compreensão teórica de propriedades importantes dos algoritmos ACO¹⁰⁴.

As características básicas do algoritmo de colônia de formigas, como explicam BONABEAU et al. 1999¹⁰⁵, são: i) a retroalimentação positiva em função das trilhas de feromônio, isto é, quanto maior o nível de feromônio melhor a qualidade da solução; ii) o feromônio virtual, cuja quantidade é acrescida nas soluções boas e decrescida em outras soluções não efetivamente boas evitando a estagnação; iii) o comportamento cooperativo das formigas, cuja exploração é coletiva; e iv) o reforço de feromônio nas trilhas de feromônio das formigas que atingiram melhores desempenhos.

Segundo Shamsipur et. al. 2007¹⁰³, o algoritmo colônia de formigas ACO é basicamente executado por um laço contendo dois procedimentos básicos: Como o problema a ser resolvido deve ser estruturado e/ou modificado e a atualização das trilhas de feromônio. Inicialmente todas as variáveis recebem a mesma quantidade de feromônio, o que representa inicialmente a mesma probabilidade de ocorrência (ou seja, $p_k=1$). O vetor

t_k (vetor feromônio) com dimensões $1 \times k$, onde k é a quantidade de variáveis na matriz (X) , recebe em todos os elementos o valor 1. A probabilidade de uma variável ser selecionada será dada pela **Equação 11** ¹⁰³:

$$P_k = \frac{t_k}{\sum_{k=1}^c t_k} \quad (11)$$

A partir deste vetor P_k uma colônia de formiga é construída, de modo que cada formiga é um vetor linha, baseada na codificação binária (0 e 1) nas variáveis. O número sendo codificado como 1, se a variável foi selecionada pela formiga e 0, se não foi. O número de variáveis em uma formiga codificada como 1 pode ser variado de 1 a c , em que c é o número total de variáveis ¹⁰³.

A quantidade de feromônio em cada preditor é variada através de procedimento iterativo, quanto maior a quantidade de feromônio depositado em uma variável maior as chances de ser selecionada por uma formiga. Então a cada iteração do algoritmo a probabilidade cumulativa de cada variável é armazenada no vetor cp_k para uma colônia com n formigas, é calculada pela **Equação 12** ¹⁰³:

$$cp_k = \sum_{i=1}^k P \quad (12)$$

O número de variáveis selecionadas por cada formiga é a solução do problema, e sua aptidão é avaliada por uma função de custo ou objeto (RMSECV)⁴⁰. Após o procedimento de avaliação, a melhor formiga deposita uma quantidade atualizada de feromônio sobre suas variáveis. Essas variáveis são as possíveis candidatas a ser a resposta final do processo de otimização^{40,103}. Ao fim de cada interação é realizada a etapa de evaporação. Isto provoca uma mais rápida convergência para o melhor

subconjunto selecionado. O algoritmo é encerrado após um número pré-definido de ciclos [40,103](#).

2.3.4.1 Deposição e evaporação dos feromônios

A deposição e evaporação dos feromônios é uma etapa importante no ajuste do ACO, de tal forma que evite a rápida convergência de todas as formigas a um caminho subótimo. Quanto a escolha dos valores iniciais e de acréscimo de feromônios, [98-101](#) relata que um decréscimo na intensidade dos feromônios favorece a exploração de diferentes caminhos, ao longo do processo de busca. E quanto à taxa de evaporação destes, [98-100](#) dizem que favorecem o esquecimento de erros ou de escolhas pobres, realizadas anteriormente, e permite uma melhoria contínua.

Por fim, o equilíbrio dos parâmetros, é a combinação do reforço positivo (depósito de feromônio) com o negativo (evaporação) conforme demonstrada na [Equação 13](#), permite que se evite, na maioria dos casos, uma convergência prematura do algoritmo para soluções, possivelmente não ruins, mas talvez, longe da ótima [106](#).

$$\tau_{ij} = (1 - \rho) * \tau_{ij}(t) + \sum_{k=1}^n \Delta \tau_{ij}^k \quad (15)$$

Onde τ_{ij} é o vetor feromônio, ρ é a taxa de evaporação dos feromônio e $\Delta \tau_{ij}^k$ é a quantidade de feromônio depositado.

O comportamento cooperativo é outro conceito importante na otimização: algoritmos de colônia de formigas fazem uso da exploração simultânea de diferentes soluções por meio de diferentes formigas [106](#). As melhores formigas influenciam a exploração das demais, através das estratégias empregadas para atualizar o feromônio nos caminhos [106](#).

2.3.4.2 Atualização da colônia de formigas

A atualização de colônia é realizada quando o procedimento de deposição e evaporação dos feromônios, realizado pelas formigas sobre as variáveis passa por um processo de atualização de feromônio. Assim a procura das variáveis com maior quantidade tem maiores chances de serem selecionadas pelas novas formigas. E a partir dessa quantidade atualizada, tais variáveis são possíveis candidatas a ser a resposta final do processo de otimização.

Em Allegrini e Olivieri ¹⁰⁷ foi proposto um novo algoritmo de seleção de variáveis com base na otimização de colônias de formigas. Nesse trabalho é relatado que a etapa de atualização da colônia é realizada por um método de seleção de roleta baseado no valor de aptidão que é atribuído a todas as variáveis possíveis e que está associado a uma probabilidade de seleção. Isto poderia ser imaginado semelhante a uma roleta num casino: uma parte da roda é atribuído a cada um dos possíveis candidatos com base nos seus valores de aptidão. Em seguida, uma seleção aleatória é feita da mesma forma quando a roda da roleta é girada. A aptidão de cada variável é dada pela quantidade de feromônio, que proporciona uma probabilidade de seleção de uma determinada variável. A seleção começa novamente seguindo o mesmo esquema de roleta, até que todas as variáveis sejam selecionadas.

Os pesquisadores observaram que no primeiro passo de tempo, todas as variáveis têm a mesma probabilidade de serem selecionadas, mas como as quantidades de feromônio são atualizadas em intervalos de tempo sucessivos, essas probabilidades serão diferentes.

2.3.5 Aplicações da Técnica de Otimização de Colônia de Formigas no campo da química

Embora bastante conhecido em problemas de otimização, este procedimento foi introduzido no campo da química há pouco mais de um década, e apenas algumas publicações baseadas em ACO foram relatadas

na literatura, por exemplo: seleção de recursos em análise QSAR ^{108,109} e análise de fluorescência ¹¹⁰.

Em Shamsipur et. al. 2007 ¹⁰³ foi aplicado pela primeira vez o algoritmo de colônia de formigas como uma nova técnica para seleção de variáveis em calibração multivariada, por mínimos quadrados clássicos (CLS), inversas mínimos quadrados (ILS) e mínimos quadrados parciais (PLS), em 4 diferentes conjuntos de dados experimentais das espectrometrias UV- Vis e NIR. As aplicações são:

- Determinação simultânea antazolina e a nafazolina em uma amostra de colírio comercial usando uma estratégia de seleção de comprimento de onda para encontrar a melhor região espectral do UV para cada fármaco separadamente. Os conjuntos de treinamento e teste utilizados foram de 15 e 36 espectros das amostras respectivamente. Um planejamento fatorial foi realizado para a escolha dos melhores parâmetros a serem incluídas nos modelos verificando o menor desvio padrão de PRESS das melhores formigas de cada execução do ACO (CLS, ILS e PLS). O desempenho do modelo foi analisado pelos valores do RMSEP de 0.6352, 0.6005 e 0.2857 para a antazolina e 0.8007, 1.3510 e 0.7455 para nafazolina, respectivamente.
- Para verificar a validade do ACO proposto para seleção do melhor subconjunto de variáveis para a determinação do índice de octano nas amostras de gasolinas registradas na região do infravermelho próximo-NIR, um conjunto de dados, 60 espectros das amostras, foi dividido em 40 para o conjunto de treinamento e 20 para o de teste. Dentre os métodos de calibração usados, a capacidade de previsão vai diminuindo na ordem ACO-PLS > ACO-ILS > ACO-CLS, onde os valores de RMSEP respectivamente são: 0.0392, 0.6745 e 1.5778.
- O desempenho do ACO para a análise de umidade e proteína nos grãos de trigo registrados por reflectância difusa NIR, mostrou-se melhor para o método dos mínimos quadrados parciais (PLS) quando comparados aos outros dois métodos. Os valores de RMSEP para os diferentes modelos multivariados indicam a superioridade da ACO-PLS sobre o ACO-ILS e o

ACO-CLS nos dois componentes analisados. A capacidade de previsão foi diminuindo na ordem ACO-PLS>ACO-CLS>ACO-ILS onde os valores de RMSEP respectivamente são: 0.0627, 1.5198 e 12.4480 para umidade e 0.2422, 0.5875 e 6.3776 pra proteína.

- Determinação simultânea de fenol e seus isômeros mononitro (o-nitrofenol, m-nitrofenol e p-nitrofenol) usando o algoritmo ACO combinado com CLS, PLS e ILS. Os números de factores adequados para a determinação dos 4 componentes (fenol, o-nitrofenol, m-nitrofenol e p-nitrofenol) foram respectivamente 5, 6, 6 e 7. O desempenho foi analisado pelos valores do RMSEP dos três métodos (ACO-CLS, ACO-PLS E ACO-ILS) para os quatro componentes respectivamente: 0.1361, 0.1436 e 0.2359 para o fenol, 4.1338, 4.3850 e 5.8914 para o-nitrofenol, 0.0741, 0.0328 e 0.5706 para m-nitrofenol e 0.0931, 0.1376 e 0.1320 para p-nitrofenol. O-nitrofenol mostrou o mais elevado RMSSEP devido a uma forte sobreposição com outros componentes. Uma comparação entre os resultados obtidos com ACO-ILS e CLS (como procedimentos de modelação simples) dão resultados comparáveis com aqueles obtidos pelo método PLS.

Allegrini e Olivieri ¹⁰⁷ propõem novo algoritmo de seleção de variáveis baseado na otimização colônia de formigas (ACO) usando na validação o procedimento de Monte Carlos e paralelização dos cálculos de regressão. O desempenho foi testado em um sistema simulado, onde se mostra uma superioridade significativa em relação aos outros métodos de seleção, tais como o algoritmo genético. Alguns conjuntos de dados espectroscópicos de infravermelho próximo foram submetidos ao algoritmo ACO-PLS como demonstrado abaixo.

- As amostras de gasolina foram submetidas ao algoritmo ACO-PLS para a determinação do índice de octano. Um conjunto de dados de 181 amostras foi dividido em 91 para calibração, 45 para validação e 45 para teste. O desempenho da previsão foi analisado pelos valores do RMSEP e comparados com o do algoritmo GA-PLS. Os dois modelos apresentaram praticamente os mesmos valores de RMSEP (0.39 para o GA e 0.30 para o ACO) e diferentes variáveis selecionadas (12 para GA e 5 para ACO). O

modelo ACO foi considerando o mais parcimonioso por ter selecionado um menor número de variáveis.

- O algoritmo ACO-PLS foi testado utilizando um conjunto de dados de 222 amostras de combustível diesel disponível na internet <http://www.eigenvector.com/data/SWRI/> para a previsão da viscosidade. Das 222 amostras, 116 foram selecionadas aleatoriamente para calibração, 58 para validação e 58 para teste. O desempenho da previsão foi analisado pelos valores do RMSEP e comparados com o algoritmo GA-PLS. Os dois modelos apresentaram os seguintes resultados de RMSEP, variáveis selecionadas e R^2 : de 0.11, 8 e 0.8912 para o GA e 0.087, 12 e 0.9337 para o ACO. Diante dos resultados, o método proposto pode ser considerado adequado para a determinação deste parâmetro.

- Xiaowei et. al. ¹¹¹ utilizaram a espectroscopia no infravermelho próximo para quantificar o teor total de antocianinas nos chás de seis tipos flores frescas. Para essa finalidade, os autores aplicaram modelos de calibração via regressão iPLS com prévia seleção de variáveis pelo algoritmos de colônias de formigas (ACO). O objetivo era avaliar o potencial do algoritmo ACO combinado com o intervalo PLS para selecionar intervalos de espectro eficiente para a quantificação de antocianinas na região do NIR. Para efeito de comparação, o modelo de calibração multivariado PLS aplicado ao espectro completo e os mínimos quadrados parciais do intervalo do algoritmo genético (GA-iPLS) com seleção de variáveis pelo algoritmo genético (GA) também foram utilizadas. O modelo ACO-iPLS ideal para o conteúdo total de antocianinas ($R = 0,9856$, $RMSECV = 0,1198$ mg / g) apresentou melhor desempenho que os modelos PLS, iPLS e GA-iPLS. Segundo os autores, a espectroscopia NIR tem um potencial significativo na determinação não destrutiva do conteúdo total de antocianinas nos chás das flores para os três modelos.

Uma outra combinação de algoritmos foi proposta por Zhang et. al, entre o algoritmo das formigas (ACO) e o algoritmo de informação mútua (MI) para seleção de variáveis a partir de dados de espectroscopia de infravermelho próximo para predição do teor total de nitrogênio em

amostras de solo. A partir das variáveis selecionadas foram construídos e avaliados modelos de regressão linear múltipla (MLR) e de regressão de vetores de máquinas de apoio (SVM), posteriormente a partir das informações espectrais completas também foram avaliados modelos de mínimos quadrados parciais (PLS). Como resultado deste estudo, os modelos que empregaram o conjunto de variáveis selecionadas pelo algoritmo obtiveram maior precisão do que os modelos PLS, além disso as variáveis selecionadas pelo ACO-MI apresentaram relação direta com o teor de nitrogênio total (NT), conferindo aos modelos alta capacidade preditiva na previsão do conteúdo de NT no solo¹¹².

O algoritmo colônia de formigas vem sendo explorado no contexto de calibração multivariada como um método de seleção de variáveis. Onde maioria dos métodos de seleção de variáveis foram desenvolvidos e empregados nesse contexto e adaptados para classificação multivariada.

Na literatura não foi encontrado nenhum trabalho que relatasse a aplicação do algoritmo ACO como um método de seleção de variáveis no contexto de classificação. Diante disso, partiu a ideia do desenvolvimento do algoritmo colônia de formigas para a seleção de variáveis em modelagem de classificação LDA.

2.4 Parâmetros de validação dos modelos de classificação

Ao construir os modelos de classificação existe a necessidade de verificar a performance da sua habilidade preditiva, no qual é baseada em função do número de acertos. Normalmente, diversos autores recorrem ao uso de parâmetros como: a Taxa de Classificação Correta¹¹³, Sensibilidade e Especificidade^{113, 114} para a validação de qualquer procedimento de modelagem. Os parâmetros mencionados são ferramentas de diagnóstico para verificar o ajuste do modelo nas etapas de validação e predição. Assim, o modelo pode ser considerado ajustado quando apresenta um percentual de classificação elevado de sensibilidade, especificidade e taxa de

classificação correta. Todos esses parâmetros podem ser determinados usando matriz de confusão¹¹⁵, como observado na **Tabela 2.1**.

Tabela 2.1: Exemplo de uma matriz de confusão

	Prevista pelo modelo	
	Classe 1 (Atribuída)	Classe 2 (Atribuída)
Classe 1 (Referência) POSITIVO	VP	FP
Classe 2 (Referência) NEGATIVO	FN	VN

VP=verdadeiro positivo; FP= falso positivo; FN=falso negativo; VN= verdadeiro negativo

As linhas da matriz de confusão correspondem à classe a qual as amostras foram rotuladas, e as colunas correspondem à classe a qual as amostras foram atribuídas pelo modelo. Para entender melhor, considere um modelo de classificação com duas classes 1 e 2 (A é classe positiva e B é a classe negativa). Assim sendo, quatro possibilidades podem ser consideradas: I) As amostras da classe A são atribuídas a classe A (verdadeiro positivo); II) as amostras da classe B são atribuídas a classe B (verdadeiro negativo); III) as amostras da classe A são atribuídas a classe B (falso negativo) e IV) as amostras da classe B são atribuídas a classe A (falso positivo) ¹¹⁶.

Com uso da matriz de confusão, a taxa de classificação correta, também conhecida como acurácia (**Equação 14**), e que pode ser conceituada como a probabilidade do modelo em avaliar a proporção de verdadeiros positivos e verdadeiros negativos na predição¹¹⁷.

$$\text{acurácia (\%)} = \frac{VP + VN}{VP + FP + FN + VN} \times 100 \quad (14)$$

Em que VP é o número de amostras que obedecem o caso I), VN é a classificação realizada no caso II), FP no caso III) e FN em IV).

A sensibilidade (**Equação 15**) mede a habilidade do modelo em detectar corretamente as amostras que pertencem à condição positiva, enquanto a especificidade (**Equação 16**) quantifica a detecção correta da condição negativa¹¹⁶.

$$Sensibilidade = \frac{VP}{VP + FP} \quad (15)$$

$$Especificidade = \frac{VN}{VN + FN} \quad (16)$$

CAPÍTULO III



ALGORITMO COLÔNIA DE FORMIGAS PARA CLASSIFICAÇÃO

3. Descrição do ACO-LDA proposto para seleção de variáveis

Nos métodos de reconhecimento de padrões supervisionados, as informações da modelagem estão contidas nos dados originários da resposta instrumental e no índice de classes para cada amostra.

Semelhante ao SPA-LDA⁴⁸, para iniciar o procedimento de seleção de variáveis o ACO-LDA também fornece como entrada: matrizes correspondentes às respostas instrumentais (conjunto de treinamento, conjunto de validação e conjunto externo para teste) e índices das classes (conjunto de treinamento, conjunto de validação e conjunto externo para teste).

O método proposto consiste numa combinação da seleção de variáveis via colônia de formiga associado a modelos de classificação baseados em análise discriminante linear, nomeadamente, ACO-LDA. O ACO acoplado ao LDA é um algoritmo iterativo que busca uma seleção de um subconjunto de variáveis de para discriminar amostras.

Conforme esquematizado no fluxograma do algoritmo ACO-LDA (**Figura 3.6**), é requerido inicialmente do usuário que o mesmo forneça alguns valores de entrada, são eles: número de formigas na colônia (**ants**), número de colônias (**c**) / interações (**int**), taxa de formigas cegas (**b**) em cada colônia e em cada ciclo, taxa de evaporação do feromônio (γ) e o número máximo de variáveis (**vmax**) que podem ser incluídas por formigas. Este valor deve sempre ser escolhido levando-se em conta os graus de liberdade de um modelo LDA. Para um problema de classificação envolvendo C classes, com I objetos de treinamentos, cujos dados serão centralizados na média das colunas, max_var assume como valor máximo I-C-1. Um outro parâmetro, denominado de min_var (correspondente ao menor número de variáveis que uma formiga pode conter) é configurado para ser igual a C.

A construção de subconjuntos de variáveis com base no critério do vetor de probabilidade de feromônio realizada pelo ACO-LDA resulta na

formação de uma colônia de formiga, de modo que cada formiga é um vetor linha, baseada na quantidade de feromônio existentes nas variáveis.

Partindo de cada variável x_j , $j = 1, \dots, J$, da Matriz $\mathbf{X}$ ($I \times J$) de treinamento, o ACO-LDA escolhe os melhores subconjuntos de variáveis de acordo com a implementação das seguintes etapas: codificação das formigas, geração da primeira colônia de formigas e do vetor feromônio, avaliação de aptidão, atualização da trilha de feromônio, atualização da colônia e parada.

3.1. Codificação das formigas

O algoritmo ACO-LDA empregou formigas binárias padrão com comprimento igual ao número de variáveis na matriz de resposta instrumental $\mathbf{X}$ ($I \times J$), onde N denota amostras descritas por variáveis J . Nesta codificação binária foram adotadas que os valores "1" e "0" indicam variáveis selecionadas e não selecionadas, respectivamente (**Figura 3.1**).

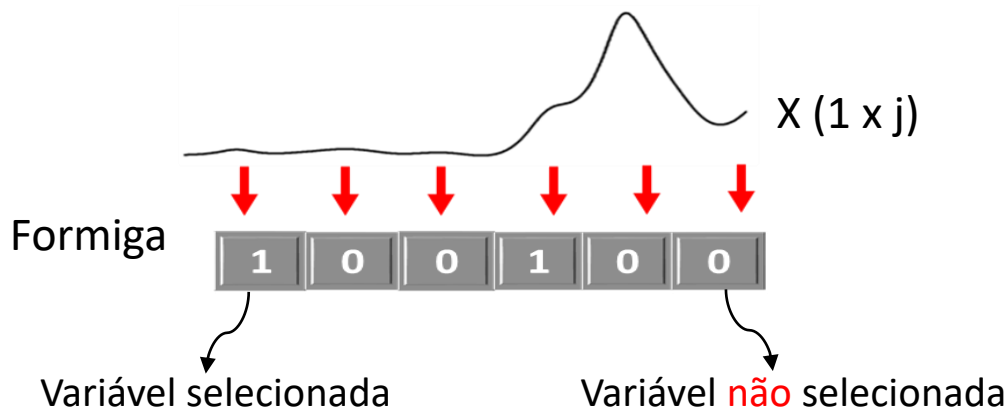


Figura 3.1. Codificação de como as variáveis são selecionadas por colônia de formigas.

Cada formiga na colônia corresponde a uma possível solução para o problema, ou seja, é um subconjunto de variáveis candidato a ser selecionado, em termos de algoritmo, cada formiga será representada por um vetor $\mathbf{a}$ de dimensões ($1 \times J$) note que J corresponde ao número de variáveis de $\mathbf{X}$ ($I \times J$). A Codificação de uma formiga inicia-se o vetor $\mathbf{a}$ (1

$\times J$) vazio, subsequentemente, um número randômico (v) entre min_var e max_var é escolhido, este número corresponde a quantidade de variáveis que a formiga em questão portará. Supondo que min_var e max_var valem 5 e 15 respectivamente, por exemplo, v pode assumir qualquer valor neste intervalo. Supondo que seja selecionado o número v igual a 7, na sequência 7 valores randômicos entre 1 e J são selecionados, estes números correspondem aos índices das variáveis selecionadas, portando estas posições em $\mathbf{a}_1$ assumem valor igual a 1 (variável selecionada), as demais posições assumem valor igual a 0 (variáveis não selecionadas).

3.2 Geração da primeira colônia de formigas e do vetor feromônio

Uma colônia inicial de formigas, a matriz **ANTS**, ($\text{ants} \times j$; onde ants e j são respectivamente o número de formigas e variáveis), é gerada aleatoriamente e armazenada em código binário (**Figura 3.2**). Além disso, é criado o vetor de feromônio $\mathbf{p}$ com a dimensão ($j \times 1$), onde todos os elementos são inicialmente iguais a 1, o que indica que todas as variáveis têm a mesma probabilidade de serem selecionadas.

$$N_{\min}=1$$

$$N_{\max}=L \text{ (quantidade máxima de variáveis)}$$

$$E_z = Z_j, j = 1, \dots, J \text{ (variáveis selecionadas pela formiga)}$$

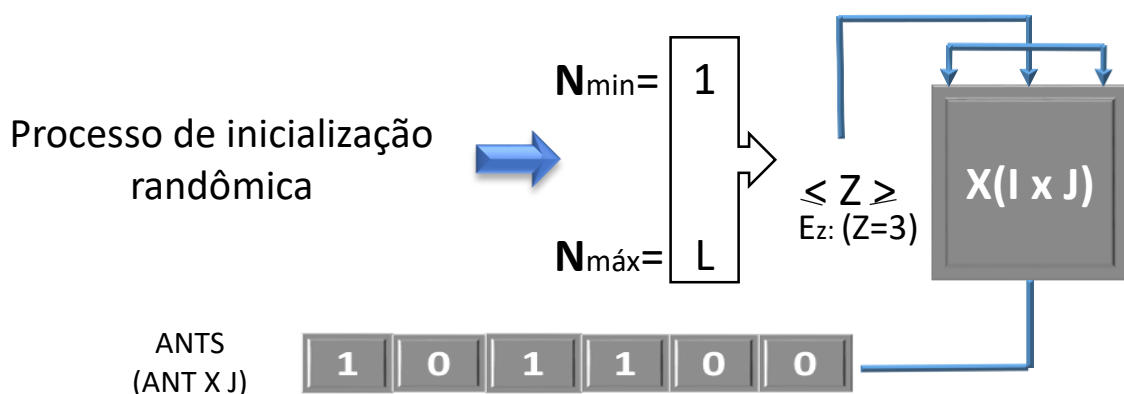


Figura 3.2. Inicialização da primeira colônia pelo processo randômico.

O procedimento descrito na **seção 3.1** é repetido **ants** vezes, gerando a quantidade de formigas inicialmente indicada pelo usuário. Os vetores **formiga****a**, são arranjados em uma matriz codificada **ANTS** (formigas × J), corresponde a primeira colônia.

3.3. Avaliação de aptidão

O modelo ACO-LDA é construído a partir de cada formiga da colônia inicial e seu poder de discriminação é avaliado. Para este propósito, uma função **G_{cost}** (**Equação 17**) foi adotada como um índice de aptidão física.

$$G_{cost} = \frac{1}{dof} \sum_{k=1}^{k_{training}} g_k \quad (17)$$

onde **G_{cost}** denota o risco médio de classificação incorreta pelo modelo ACO-LDA quando um subconjunto de variáveis (em avaliação) é empregado. O valor **dof** (**Equação 18**) é o grau de liberdade dado por:

$$dof = K_{training} - l - tc \quad (18)$$

Em que **K_{training}** é o número de amostras no conjunto de treinamento. Se este conjunto for usado para orientar a seleção de variáveis; **l** é o número de variáveis incluídas para cada formiga (**l = tc, ..., vmax**) e **tc** é o número de classes alvo. Se um conjunto de dados externo for usado para orientar a seleção de variáveis, **dof** é igual a **k_{ext} - tc** (o número de amostras no conjunto de dados externo). E **g_k** é o risco de classificação incorreta da **k**-ésima amostra de treinamento ou um objeto externo **X_k**, pertencendo ao

conjunto **K** da amostra usado para guiar a seleção de variáveis. O **g_k** é definido como:

$$g_k = \frac{r^2(\mathbf{x}_k, \boldsymbol{\mu}_{nk})}{\min_{nw \neq nk} r^2(\mathbf{x}_k, \boldsymbol{\mu}_{nw})} \quad (19)$$

onde $r^2(\mathbf{x}_k, \boldsymbol{\mu}_{nk})$ e $r^2(\mathbf{x}_k, \boldsymbol{\mu}_{nw})$ são respectivamente o quadrado das distâncias de Mahalanobis entre a **k**-ésima amostra **x_k** (um vetor linha) e o centro de sua verdadeira classe (com índice **n_k**), e o centro da classe errada mais próxima (com índice **n_w**). Os vetores linha **μ_{nk}** e **μ_{nw}** são respectivamente as médias populacionais da classe verdadeira e da classe errada mais próxima. As distâncias de Mahalanobis ao quadrado foram calculadas pela expressão:

$$r^2(\mathbf{x}_k, \boldsymbol{\mu}_{nk}) = (\mathbf{x}_k - \boldsymbol{\mu}_{nk})\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_{nk})^T \quad (20)$$

E

$$r^2(\mathbf{x}_k, \boldsymbol{\mu}_{nw}) = (\mathbf{x}_k - \boldsymbol{\mu}_{nw})\boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_{nw})^T \quad (21)$$

onde $\boldsymbol{\Sigma}^{-1}$ é o inverso da matriz de covariância, estimada de acordo com o procedimento LDA descrito no trabalho de W.Wu et.al.¹¹⁸. Idealmente, o objeto **x_k** deveria estar próximo ao centro da verdadeira classe e distante dos centros de todas as outras classes; consequentemente, de acordo com a **Equação 19**, **g_k** deve ser o menor possível.

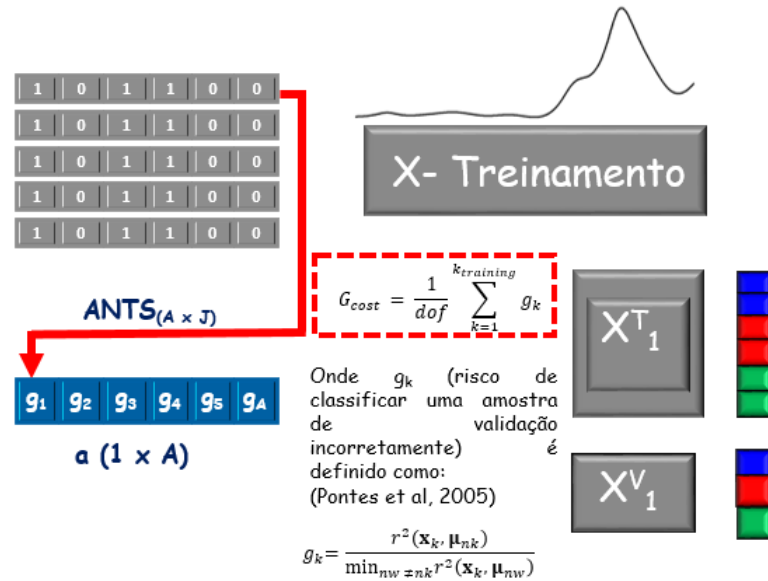


Figura 3.3. Avaliação do poder discriminante para cada modelo construído das formigas iniciais.

A cada vetor linha de formigas corresponderá um valor de G_{cost} , estes valores são armazenados em um vetor **cost** ($1 \times A$). Esses valores correspondem a uma medida da capacidade preditiva de cada subconjunto de variáveis, ou seja, é a capacidade de gerar boas soluções. Essa informação é usada para a atualização da trilha de feromônio (vetor de feromônio artificial).

3.4 Atualização da trilha de feromônio:

O vetor de feromônio artificial é atualizado em iterações sucessivas de acordo com a **Equação 22**. A trilha de feromônio é aqui representada como um vetor $\mathbf{p}$ ($J \times 1$). Note que cada elemento do vetor $\mathbf{p}$ está relacionado com a quantidade de feromônio depositada na j -ésima variável. Inicialmente, a cada variável é atribuída igual quantidade de feromônio, denotado como 1 em todas as posições do vetor $\mathbf{p}$. A atualização da quantidade de feromônio depositada em cada variável é uma função do poder discriminante (G_{cost}) do modelo LDA construído levando em conta cada vetor linha de **ANTS**.

$$p(i) = (1 - \gamma)p(i - 1) + \Delta p \quad (22)$$

Onde $p(i)$ é o vetor de feromônio atualizado ($J \times 1$) na iteração atual i ($i=1,2,\dots,c$), o termo $p(i - 1)$ é o vetor feromônio da iteração anterior, e γ é a taxa de evaporação do feromônio, cujo valor implementado foi igual a 0,65 (65%). O operador γ impede a inclusão de variáveis não informativas. Note que na primeira iteração p vale 1 para todas as variáveis.

Na **Equação 22**, Δp ($j \times 1$) é o vetor de mudanças de feromônio que corresponde ao incremento de feromônios depositados sobre as variáveis pertencentes a cada formiga na interação, conforme expresso matematicamente na **Equação 23**.

$$(\Delta p)_j[v(l)]_{ant} = -\log(G_{cost})_c \quad (23)$$

onde $v(l)$ é um elemento do vetor var ($1 \times l$) das variáveis selecionadas que correspondem a cada formiga.

A **Figura 3.4** ilustra como é realizado o processo de atualização da trilha de feromônio.

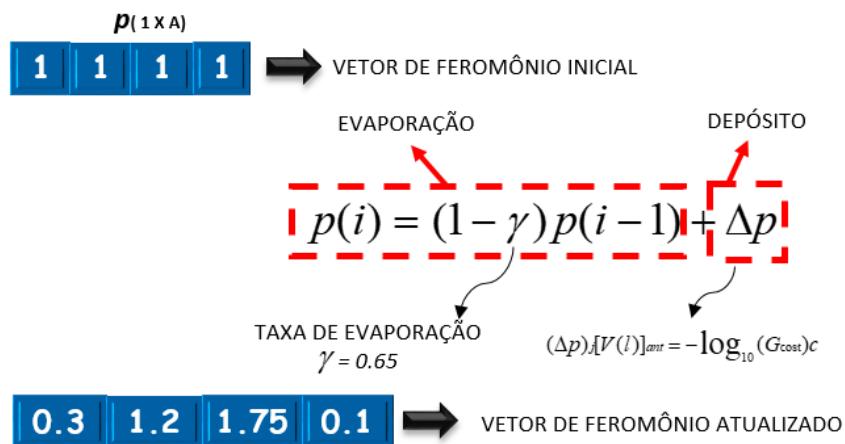


Figura 3.4. Atualização da trilha de feromônio

Após a completa atualização de ***p***, cada variável que tinha o valor 1 de feromônio, passará a ter um valor correspondente a sua capacidade de gerar modelos LDA com baixo risco de uma classificação incorreta e logo uma nova colônia de formiga é gerada.

3.5. Atualização da colônia

Após a atualização da trilha de feromônios, uma nova colônia de formigas é gerada com base em duas contribuições: ($ants_{pher}$ e $ants_{blind}$) os das formigas com feromônio e das formigas cegas. A primeira contribuição leva em consideração a quantidade de feromônio depositada em cada variável na iteração atual usando o método da roleta¹¹⁹. Assim, uma porcentagem das formigas que comporão a nova colônia, serão formadas escolhendo as variáveis de forma probabilística conforme a **Equação 24**. Quanto maior a quantidade de feromônio em uma variável, maior a probabilidade de ser selecionada para compor a nova colônia. A seleção dessa variável é proporcional à aptidão promovida por ela mesmo.

$$prob(j) = \frac{p(j)}{\sum_{j=1}^j p(j)} \quad (24)$$

Onde $p(j)$ é o j -ésimo elemento do vetor $p(j \times 1)$.

A probabilidade da variável a ser selecionada é a razão entre o conteúdo de feromônio e a soma do conteúdo de feromônio de todas as variáveis. As outras formigas da colônia são denominadas de formigas cegas. A contribuição ***ants_{blind}*** compreende um procedimento que envolve a geração aleatória de formigas cegas, o que impede um mínimo local. Finalmente, ***ants_{pher}*** mais ***ants_{blind}*** deve ser igual a ***ants*** e a taxa de

formigas cegas (**b**) em cada colônia e em cada ciclo não deve ser maior que 0,35 (35%) para evitar problemas de convergência.

A **Figura 3.5** ilustra basicamente o processo de atualização da colônia.

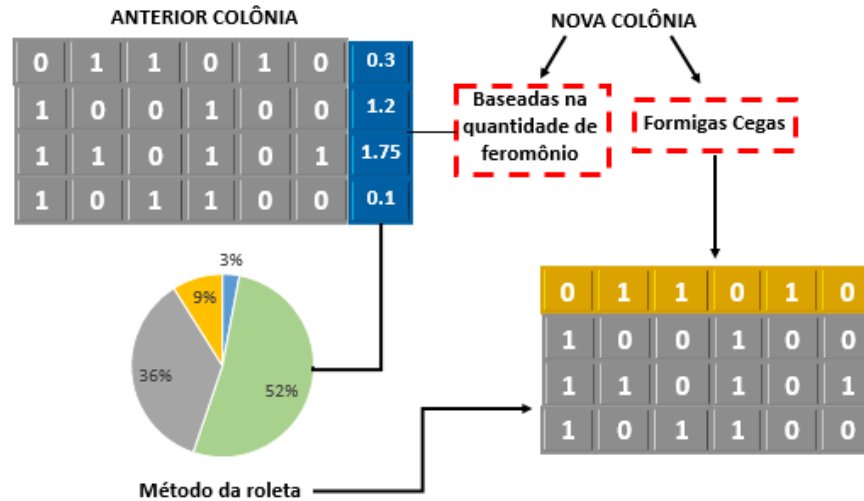


Figura 3.5. Atualização da colônia

A estratégia consiste em gerar uma proporção de novas formigas cegas (**b**) na colônia de forma randômica, como descrito na **seção 3.1**. O tamanho da colônia de formigas será igual ao valor pré-estabelecido pelo usuário (**ants**).

3.6 Critério de parada

O critério de parada, adotado para finalizar o procedimento de otimização, foi estabelecido quando número máximo (**c**) de iterações (ou colônias) é atingido, conforme mostrado na **Figura 3.6 A**.

A fim de evitar a seleção de variáveis não informativas no modelo ACO-LDA final, uma abordagem Double Monte Carlo¹²⁰(dMC) adaptada da validação cruzada dupla repetida¹²¹ foi aplicada. Tal validação dMC corresponde a ciclo aninhados, conforme ilustrado na **Figura 3.6 B**.

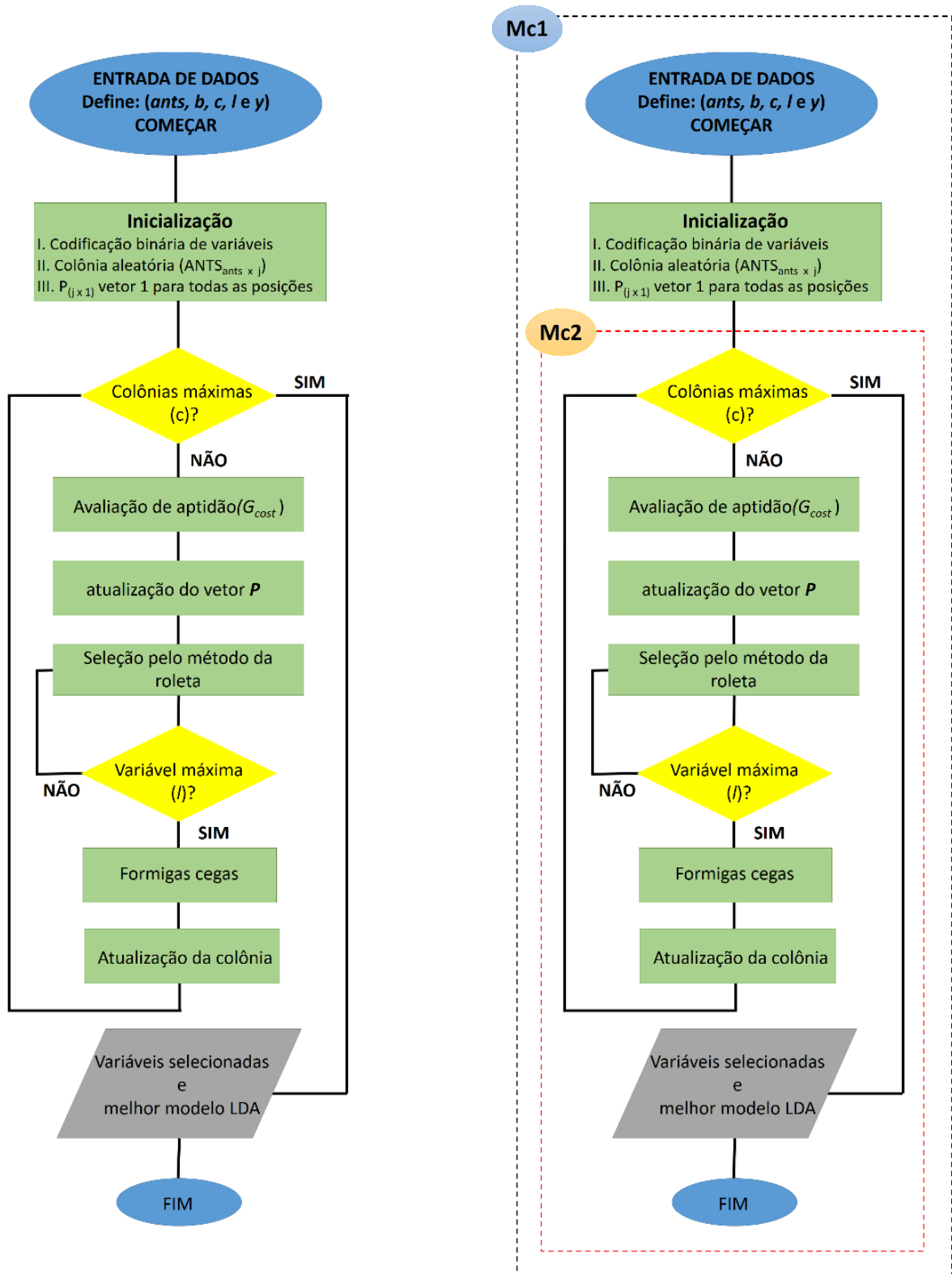


Figura 3.6: Fluxograma do algoritmo ACO-LDA. **A.** Sem Double Monte Carlo e **B.** Com Double Monte Carlo.

Na primeira etapa, ACO-LDA é executado mc_1 vezes até que o mesmo conjunto (amostras de treinamento e validação) seja selecionado aleatoriamente. Este procedimento é repetido mc_2 vezes considerando um novo conjunto de amostras. Depois disso, um histograma de frequência das variáveis selecionadas é gerado levando-se em consideração que as variáveis são ponderadas inversamente proporcionais ao valor final de G_{cost} (para dar comparativamente mais importância a melhores soluções) e o fator de inflação de variância - VIF¹²² (para evitar que as variáveis tenham multicolinearidade). O histograma é então normalizado para um valor máximo de 1. As variáveis com valores de histograma maiores que uma tolerância predefinida são empregadas para construir um modelo LDA final a ser usado para prever um conjunto independente de amostras.

CAPÍTULO IV

***ESTUDO DE CASO I e II:
CLASSIFICAÇÃO DE ÓLEOS VEGETAIS
E CHÁS***



4. Classificação de óleos vegetais e chás

4.1 Introdução

4.1.1. Óleos vegetais comestíveis

O óleo vegetal é uma gordura obtida de plantas. É extraído na maioria das vezes das sementes ou grãos destes vegetais conhecidos como oleaginosos. Os óleos destinados ao consumo humano são submetidos a um processo de refino cuja finalidade é melhorar sua aparência, sabor, odor e estabilidade por remoção de alguns componentes como: ácidos graxos livres, proteínas, corantes naturais, umidade e compostos voláteis e inorgânicos⁶⁹.

Os óleos vegetais são constituídos majoritariamente por triglicerídeos, que são compostos de três moléculas de ácido graxo (AcG) com uma molécula de glicerol. As unidades acilas são os principais constituintes dos óleos vegetais, representando cerca de 95% da massa molecular dos triglicerídeos. Em pequena quantidade aparecem mono e diglicerídeos, ácidos graxos livres, glicolipídeos, proteínas, fosfolipídeos, esteróis entre outras substâncias.¹²³

Os grãos de canola produzidos no Brasil possuem cerca de 24 a 27% de proteína e de 34 a 40% de lipídio (óleo) e elevada quantidade de ômega 3, gorduras monoinsaturadas e vitamina E^{124,125}. Óleos de girassol tem uma baixa quantidade de ácidos graxos saturados, alto teor de ácido linoléico e vitaminas, além de apresentar características organolépticas agradáveis¹³⁰. O grão de soja possui cerca de 40% de proteínas, 20% de lipídios (óleo), 5% de minerais e 34% de carboidratos. Os principais AcGs que compõem o óleo de soja são o linoléico, oléico, palmítico e linolênico¹²⁷. Os principais ácidos graxos que compõem o óleo de milho são o linoléico (até 62%) e oléico (até 42%)¹²⁸.

Como pode observar, entre os ácidos graxos que constituem os óleos vegetais os mais comuns são o oleico ($C_{18}H_{34}O_2$), o linoléico ($C_{18}H_{32}O_2$) e o linolênico ($C_{18}H_{30}O_2$). A **Figura 4.1** representa a estrutura molecular desses ácidos.

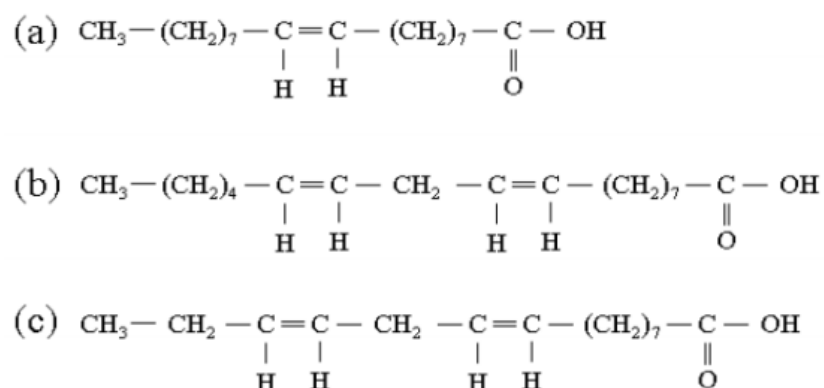


Figura 4.1: Estruturas moleculares dos ácidos (a) oleico, (b) linoléico, (c) linolênico.

Para avaliar a qualidade de um óleo vegetal comestível com técnicas espectroscópicas para classificar óleos vegetais comestíveis de acordo com tipo de oleaginosa (milho, canola, soja e girassol) foi utilizado como conjunto de dados para aplicação do algoritmo ACO-LDA.

4.1.2 Chás

O chá é uma infusão preparada a partir das folhas vaporizadas e secas do arbusto perene *Camellia sinensis*¹²⁹ e são classificados em três categorias conforme o processo de fabricação: fermentado (preto), não-fermentado (verde) e o semifermentado (oolong). As categorias utilizadas na aplicação desse trabalho foi o chá verde e preto.

O chá verde é produzido a partir de folhas frescas da planta, as quais são apenas escaldadas e fervidas, ocorrendo uma rápida inativação da enzima polifenol oxidase, o que mantém preservado seu teor de polifenóis e o torna mais rico em catequinas e compostos com atividades funcionais¹³⁰. No chá verde, estão presentes, além das catequinas, outros compostos orgânicos, tais como cafeína e aminoácidos¹³¹. Na fabricação do chá verde, as enzimas são inativadas imediatamente após a colheita das folhas. O chá preto é derivado de folhas envelhecidas pela oxidação aeróbica

das catequinas, ocorrendo uma fermentação¹³⁰. Na produção do chá preto, as catequinas são oxidadas enzimaticamente, gerando uma mistura complexa de polifenóis, constituída de teaflavinas, teasinensinas e tearubiginas^{131,132}. As catequinas oxidadas reagem com as moléculas precursoras presente no chá verde e produzem compostos voláteis no chá preto, que são dependentes principalmente de variáveis genéticas, de cultivo e de fabricação¹³³. A diferença entre o chá verde e o chá preto depende de quando as enzimas foliares são inativadas durante o processamento¹³¹.

O Desenvolvimento de estratégia para analítica para classificação dos chás de acordo com o tipo e origem geográfica pode fornecer segurança aos compradores e consumidores, além de poder prevenir a rotulagem fraudulentas de chás.

4.2. Metodologia

Nesse capítulo são descritos e avaliados os bancos de dados desenvolvidos nos trabalhos de tese de Márcio José Coelho de Pontes (óleo vegetal comestível)⁵¹ e Paulo Henrique Gonçalves Dias Diniz (chás)⁶¹ para a modelagem do algoritmo proposto (ACO-LDA) bem como dos algoritmos usados para comparação (GA-LDA e PLS-DA). E por fim, os resultados das classificações envolvendo ambos conjuntos de dados dos modelos citados são apresentados.

O conjunto de dados dos chás, está relacionado com a classificação simultânea de chás de acordo com a sua variedade e origem geográfica e o conjunto de dados dos óleos está relacionado ao tipo de oleaginosas.

4.2.1 Conjunto de dados para análise multivariada

Dados UV-Vis dos óleos vegetais comestíveis

Neste estudo de caso, foram consideradas quatro classes-alvo: Classe 1 (amostras de milho -29), Classe 2 (amostras de soja -30), Classe

3 (canola - 29 amostras) e Classe 4 (girassol -31 amostras). Os espectros UV-Vis foram o espectrofotômetro de absorção molecular UV-VIS HP 8453, modelo Hewlett Packard, de uma cubeta de fluxo de quartzo com trajeto óptico de 1 mm. O espectro foi registrado na região de 220 nm a 400 nm, com resolução de 1 nm. Para todo o caso, as amostras de óleo são previamente diluídas com álcool n-butílico (99% p / p) na proporção de 1: 300 (v / v).

Dados NIR dos Chás

Foram empregadas 100 amostras de chás de diferentes variedades (preto e verde) e origens geográficas (Brasil, Argentina e Sri Lanka). Cinco classes-alvo são consideradas, nomeadamente: chás verdes argentinos (classe 1 -AG); Chás argentinos pretos (classe 2 - AB); Chá verde brasileiro (classe 3- BG); Chás brasileiras pretas (classe 4 - BB) e chás pretos do Sri Lanka (classe 5 SB). Todas as classes continham vinte amostras de diferentes marcas. Os espectros NIR, em duplicata, foram registados na faixa de 715 a 2700 nm, utilizando um espectrofotômetro FTIR FTLA2000-160 (Bomem®) a uma temperatura ambiente de 23 ± 1 ° C. O espectro de uma pastilha de politetrafluoroetileno (PTFE) foi utilizado como branco. Após a exclusão das regiões de ruído excessivo e saturação do detector, o intervalo de trabalho considerado em todos os cálculos subseqüentes foi de 1001 a 2500 nm.

4.2.2 Procedimentos quimiométricos

Todos os cálculos foram realizados no ambiente MatLab 2012b. O Classification Toolbox 5.2¹³⁴ para Matlab foi usado na modelagem PLS-DA. Nenhum pré-processamento foi realizado para os espectros de UV-Vis ou NIR.

Para modelagem, os espectros brutos das amostras de óleo comestível e chá foram divididos em conjuntos de Xdata (72 e 75) e de teste (47 e 25) usando o algoritmo clássico de Kennard-Stone (KS) ¹³⁵

respectivamente, para evitar problemas de extrapolação dos modelos na fase de previsão. O conjunto X_{data} foi usado para construir e validar os modelos, e incluiu a seleção de variáveis para ACO-LDA e GA-LDA. A cada ciclo de mc_1 , o conjunto de dados de treinamento-validação foi dividido aleatoriamente em X_1 (modelo) e X_2 (validação), sempre mantendo a proporção de 70% (X_1) e 30% (X_2). As amostras de teste foram utilizadas apenas para comparação e avaliação final dos modelos de classificação. O conjunto de teste não participou da etapa da construção do modelo e foi usado com a finalidade de comparar o desempenho dos diferentes modelos em função da taxa de classificação correta.

O algoritmo ACO-LDA foi executado usando os seguintes parâmetros de entrada: quantidade de formigas (*ants*) em colônia: 50; número de colônias (*c*): 100; taxa de formigas cegas (*b*) em cada colônia e em cada ciclo: 0,35; taxa de evaporação de feromônio : 0,65; e número máximo de variáveis (*l*) que podem ser selecionadas por uma formiga. A função custo empregada foi a minimização do G_{cost} no conjunto de validações.

O algoritmo GA-LDA¹³⁶ também empregou codificação binária (um gene "1" ou "0" indica uma variável selecionada e não selecionada, respectivamente). Cromossomos com comprimento igual ao número de comprimentos de onda no espectro foram utilizados. A probabilidade de um determinado cromossomo ser selecionado para o acasalamento foi proporcional à sua adequação (dada pelo inverso da função custo, G_{cost}) de acordo com o método da roleta¹³⁷. Operadores genéticos de cruzamento e mutação foram utilizados com probabilidades de 60% e 10%, respectivamente. O tamanho da população foi mantido fixo em 50 indivíduos que evoluíram por 100 gerações. Cada geração foi completamente substituída por seus descendentes. Em ambos os casos, o número de ciclos mc_1 e mc_2 foi igual a 150 e 150, respectivamente.

Para o modelo de análise discriminante por mínimos quadrados parciais (PLS-DA), Monte Carlo foi usado para seleção de variáveis latentes com 300 ciclos para os dois conjuntos de dados. Os resultados da validação

expressam valores médios em todos os ciclos MC. O valor limiar adotado foi de 0,5. Quando um valor acima de 0,5 é previsto, a amostra é considerada como pertencente á classe em estudo, enquanto um valor abaixo de 0,5 indica que a amostra não pertence aquela classe.

4.3. Resultados e discussão

4.3.1. Classificação de óleo comestível em relação ao tipo de oleaginosa.

4.3.1.1. Perfil espectral das amostras de óleo comestível

Os espectros de absorção molecular dos quatro tipos de óleos vegetais comestíveis (canola, girassol, milho e soja), analisados na região de 220 nm a 400 nm, são mostrados na **Figura 4.2**.

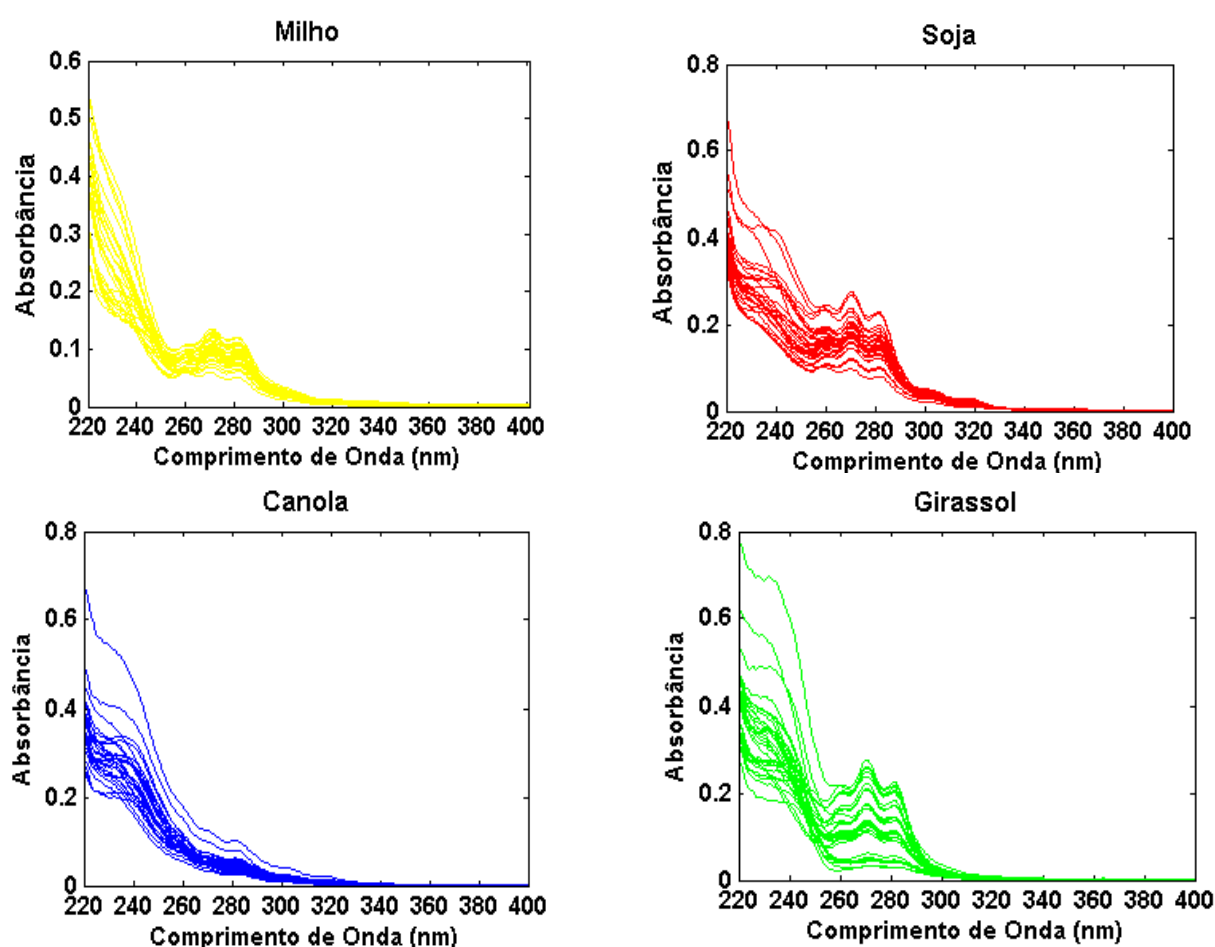


Figura 4.2. Espectros UV-VIS das 4 classes das amostras de óleos vegetais comestíveis analisados.

As absorções das amostras ocorrem na faixa do UV como observado na **Figura 4.2**. Os quatro tipos de óleos apresentam um pico de absorção abaixo de 1.0 (Abs <1.0), demonstrando que o principal composto capaz de gerar pico de absorção nessa região é o triacilglicerol, que compõe os óleos. Observa-se que os óleos de milho, soja e girassol mostram bandas de absorção bem caracterizadas, enquanto no óleo de canola não é tão aparente. Podemos analisar melhor nos espectros médios das classes (**Figura 4.3**).

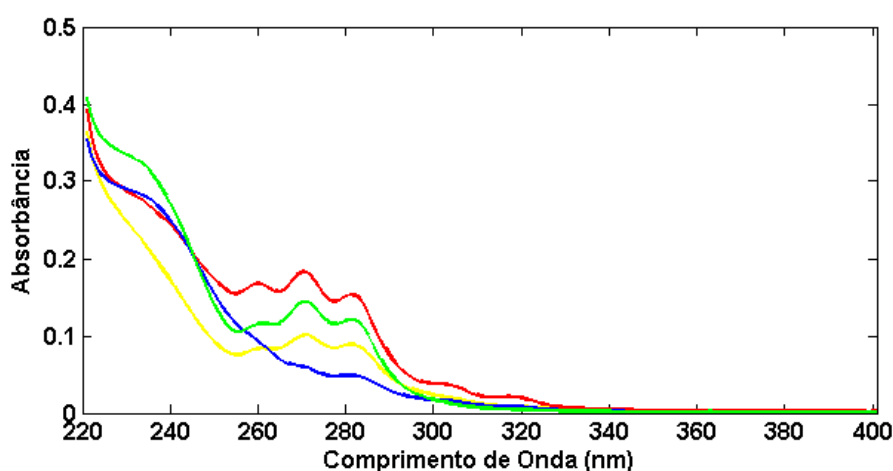


Figura 4.3. Espectros médios de canola (linha azul), soja (linha verde), milho (linha amarela) e girassol (linha vermelha)

Observando os espectro médio das classes, podemos ver uma nítida separação entre as classes. Mas quando analisamos o espectro com todas as mostras completas, vemos que alguns possuem perfis semelhantes (**Figura 4.4**). Isto está relacionado à grande semelhança na composição dos óleos analisados, torna-se impossível a distinção entre os diferentes tipos de óleos vegetais por meio da inspeção visual dos espectros. Para isto, tornou-se indispensável o uso de ferramentas quimiométricas.

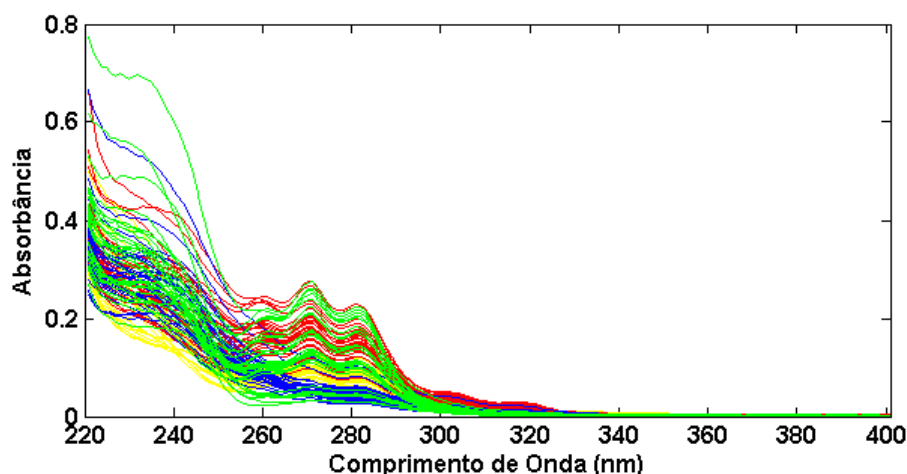


Figura 4.4. Espectros UV-VIS de todas as 4 categorias das amostras de óleos vegetais comestíveis analisados.

As bandas que aparecem para o óleo vegetal são devido às duplas ligações isoladas carbono-carbono presentes nos ácidos graxos, revelando transições do tipo $\pi \rightarrow \pi^*$ correspondentes a componentes ou grupos cromóforos¹³⁸. As bandas de absorção para os óleos de soja, milho e girassol são mais nítidas nas região entre 260, 270 e 280 nm⁵¹, enquanto para os óleos de canola apenas a banda em 290 nm⁵¹ é mais evidente. As bandas de absorção na região de 230 a 270 nm podem estar associadas a instauração conjugada dos ácidos graxos¹³⁹. E a banda intensa em torno de 232 nm pode estar associada à absorção de dienos conjugados¹⁴⁰ e as demais bandas podem estar associadas a absorção de alguns ácidos graxos que absorvem em regiões abaixo de 375 nm, tais como ácido oleico¹⁴¹. Com isso pode-se inferir maior conteúdo de dienos conjugados no óleo de girassol, visto que, nessa região obteve uma absorção maior dos demais. A seguir o desempenho do ACO-LDA para classificação destes quatro tipos de óleos é apresentado, bem como o do GA-LDA e PLS-DA.

4.3.1.2. ACO-LDA

Os melhores subconjuntos de variáveis selecionadas são indicados na **Figuras 4.5**. O modelo ACO-LDA apresentou como solução ótima, com

um subconjunto contendo sete variáveis (comprimento de onda), para um valor limite ótimo igual a 0,35.

As barras azuis correspondem à frequência de seleção das variáveis nos ciclos repetidos de Monte Carlo, visando avaliar a convergência do algoritmo para as mesmas soluções. Pode-se observar que a maior frequência encontra-se em torno de 300nm. A linha horizontal representa um valor limite otimizado em relação à taxa de erros de validação cruzada. O valor limite corresponde à probabilidade de selecionar uma variável específica em uma única execução de algoritmo ACO para classificação de óleos comestíveis.

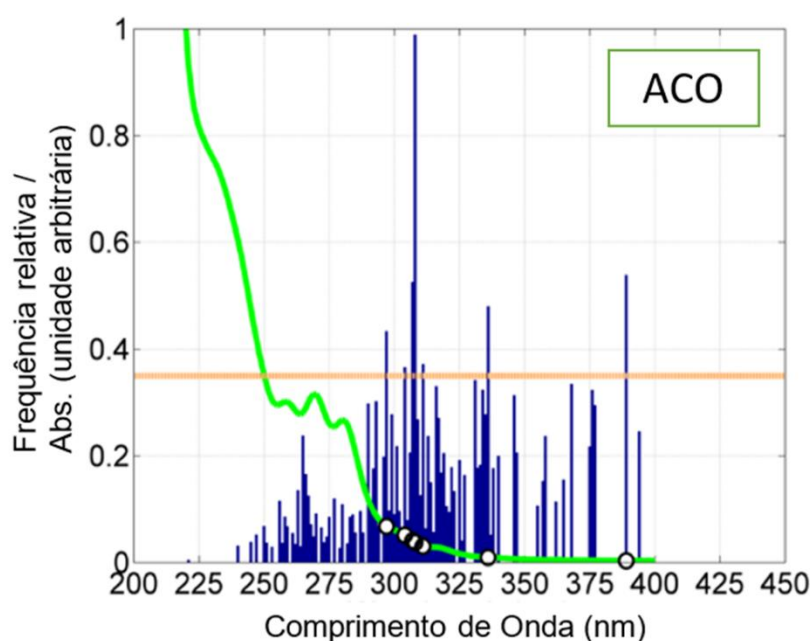


Figura 4.5: Comprimento de onda ou variáveis latentes (círculos brancos) selecionados pelos algoritmos ACO-LDA dos dados do UV-Vis.

Como se pode notar, dos sete comprimentos de onda (297, 304, 307, 308, 311, 336, 389 nm) selecionados pelo ACO-LDA, cinco estão associados a pontos característicos dos espectros das 4 variedades de oleaginosas que portam informações suficientes para distingui-las e uma em regiões de baixa intensidade de absorção.

A **Figura 4.6** representa os valores do fator de inflação de variância (VIF) para todas as variáveis. As marcadas por um círculo preto são as

variáveis selecionadas pelo ACO-LDA. Observe que o VIF para todas variáveis, selecionadas pelo ACO, possuem valores menores que 4, o que pode levar ao acerto. Isso sugere que o ACO-LDA pode estar selecionando variáveis uteis, o que aumenta a estabilidade e a propagação de acertos do modelo.

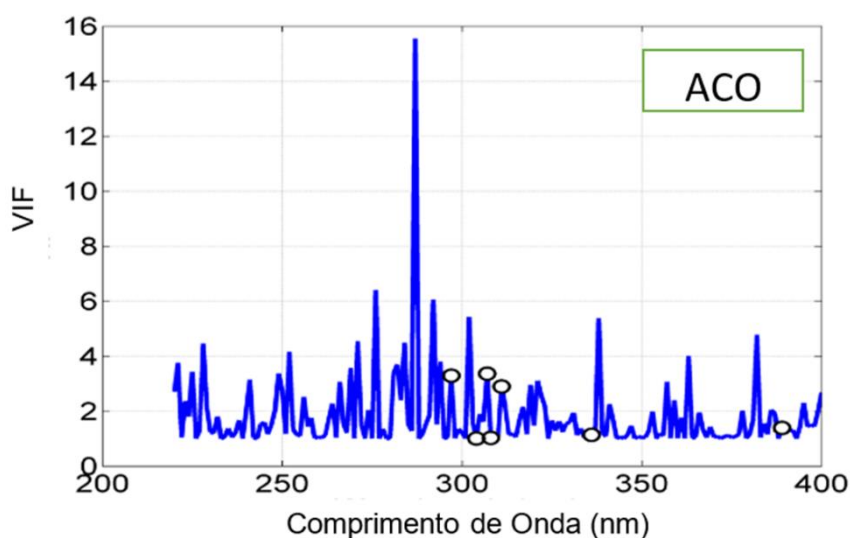


Figura 4.6: Valores VIF (linha azul sólida) para as variáveis selecionadas (círculos brancos com borda preta) por ACO-LDA dos dados do UV-Vis.

O modelo LDA obtido com as sete variáveis selecionadas pelo ACO foi, então, aplicado à classificação de amostras do conjunto de teste. Como resultado, todas as amostras foram classificadas corretamente. O gráfico dos escores obtidos pela LDA utilizando os sete comprimentos de onda selecionados pelo ACO são apresentados na **Figura 4.7 A**.

Os gráficos dos escores dos modelos ACO-LDA, revelam informações úteis. Onde pode-se observar que a FD1 é responsável por separar de fato as classes de acordo com o tipo de óleo. Ou seja, nenhum erro foi obtido durante a construção do modelo, assim como na previsão dos conjuntos externos de amostras na classificação de todas as oleaginosas deste conjunto.

Pode-se observar a eficiência do ACO-LDA em selecionar variáveis espectrais que sejam suficientemente discriminantes para o estudo de classificação de óleos vegetais comestíveis.

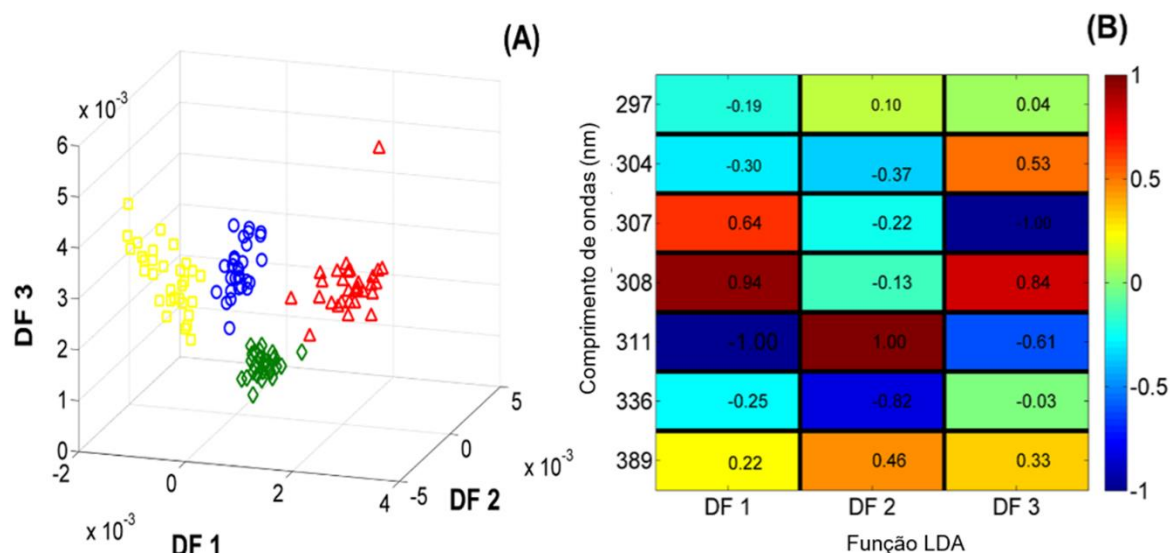


Figura 4.7: Gráficos de escores e mapa térmico dos pesos do modelo ACO-LDA (A e B), para classificação de óleos comestíveis. Círculo azul (canola), triângulo vermelho (soja), quadrado amarelo (milho) e diamante verde (girassol).

A partir dos pesos na **Figura 4.7. B**, observa-se que entre as variáveis escolhidas pelo modelo ACO-LDA, as variáveis de maiores pesos para FD1 foram os comprimentos de ondas de 307, 308 e 311 nm, para o FD2 foram 311 e 336 nm e para o FD3 foram as mesmas variáveis que FD1. Logo, a função discriminante que obtiveram a melhor discriminação das classes foram FD1 e FD2. Esses comportamentos dos óleos podem estar relacionados aos diferentes ácidos graxos, os quais diferem entre si pelo tamanho da cadeia carbônica, número e orientação das ligações duplas. Nesses óleos, a quantidade de ácidos graxos oleico (18 carbonos, 2 insaturação) e linoleico (18 carbonos, 2 insaturação) são os que podem justificar a discriminação das oleaginosas.

4.3.1.3. GA-LDA

O algoritmo GA, para o valor t ideal "0,45", selecionou quinze variáveis espectrais para a construção do modelo GA-LDA: 299, 305, 306, 307, 316, 317, 321, 325, 337, 338, 340, 370, 377, 380, 382 nm indicados na **Figura 4.8**. Dentre dessas variáveis, algumas são localizadas em regiões onde nenhuma informação é evidenciada.

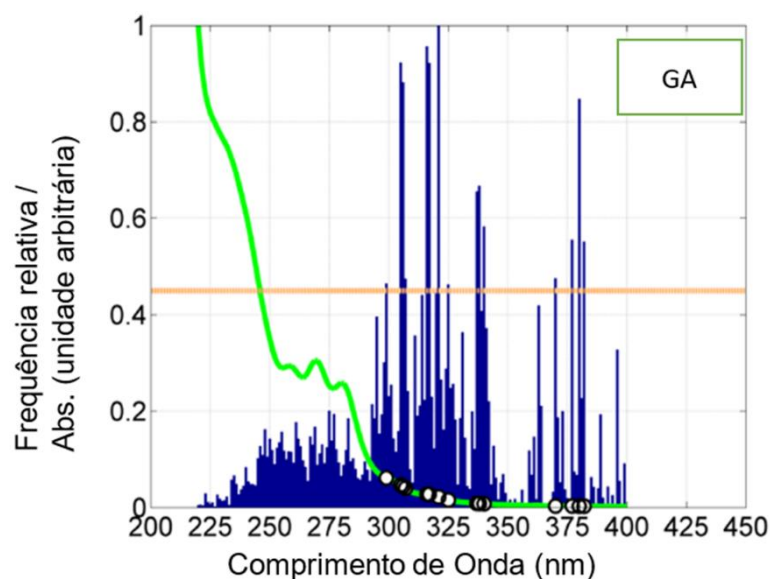


Figura 4.8: Comprimento de onda ou variáveis latentes (círculos brancos) selecionados pelos algoritmos GA-LDA dos dados do UV-Vis.

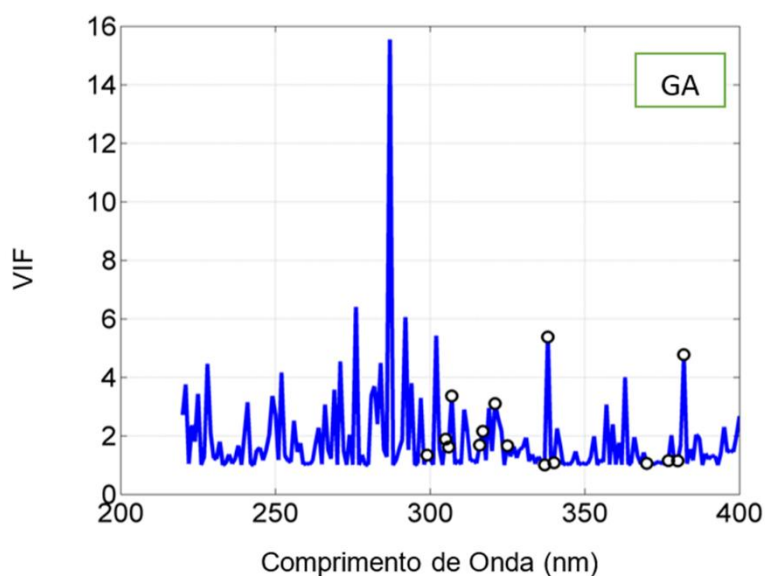


Figura 4.9: Valores VIF (linha azul sólida) para as variáveis selecionadas (círculos brancos com borda preta) por GA-LDA dos dados do UV-Vis.

A **Figura 4.9** representa os valores de VIF para todas as variáveis, como já mencionado anteriormente no ACO-LDA, agora marcadas por um círculo branco com bordas pretas são variáveis selecionadas por GA-LDA.

Observe que o VIF para algumas variáveis, selecionadas pelo GA, possuem valores maiores que 4, o que pode levar em alguns casos à propagação de erros. Isso sugere que o GA-LDA pode estar selecionando variáveis redundantes, o que aumenta a instabilidade e a propagação do erro do modelo.

As quinze variáveis selecionadas foram as melhores para a construção do modelo ACO-LDA e então aplicado na classificação das amostras do conjunto de previsão (amostras externa). A **Figura 4.10** mostra o gráfico dos escores de todas as amostras envolvidas no conjunto de treinamento e teste analisadas nesse estudo. Neste gráfico observa-se a formação dos quatro grupos de amostras correspondentes as quatro oleaginosas analisadas. As classes analisadas para esse conjunto estão bem distribuídos e agrupados cada uma em seu grupo de oleaginosas, existindo dentre elas uma sobreposição de duas amostras do conjunto teste em outro grupo que não o seu. Isso provavelmente está relacionada as composições química entre as amostras de girassol e milho.

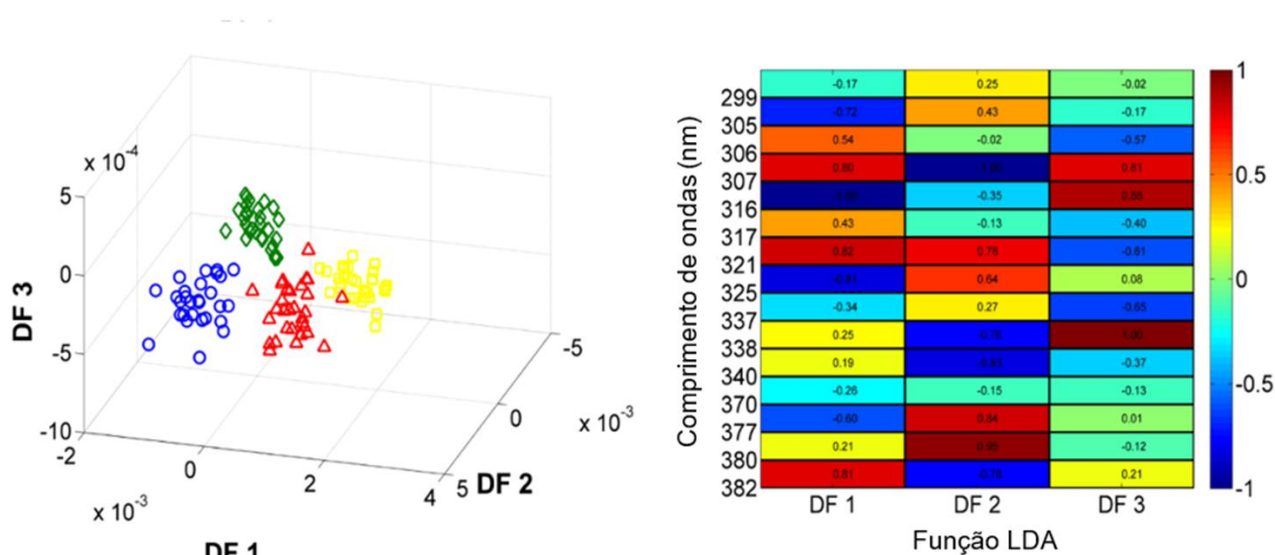


Figura 4.10: Gráficos de escores e mapa térmico dos pesos do modelo GA-LDA, para classificação de óleos comestíveis. Círculo azul (canola), triângulo vermelho (soja), quadrado amarelo (milho) e diamante verde (girassol).

Observado o gráfico dos pesos (**Figura 4.10**), as variáveis selecionadas pelo GA que melhor contribuem para a separação das classes em FD1 foram os comprimentos de onda: 305, 307, 316, 321, 382 nm; para FD2 foram de 307, 321, 340, 370, 377, 380, 382 nm e para o FD3 foram de 307, 316, 338 nm. Da mesma forma como o ACO, foram selecionados variáveis que tivessem pesos discriminantes entre as classes.

4.3.1.4 PLS-DA

Na construção do modelo PLS-DA, a quantidade de variáveis latentes são aquelas que obtiverem o menor número de erros e maior frequência no processo de Monte Carlo. Na figura abaixo (**Figura 4.11**) uma quantidade de 7 variáveis latentes foram necessárias para obtenção dos resultados de classificação das amostras de óleos.

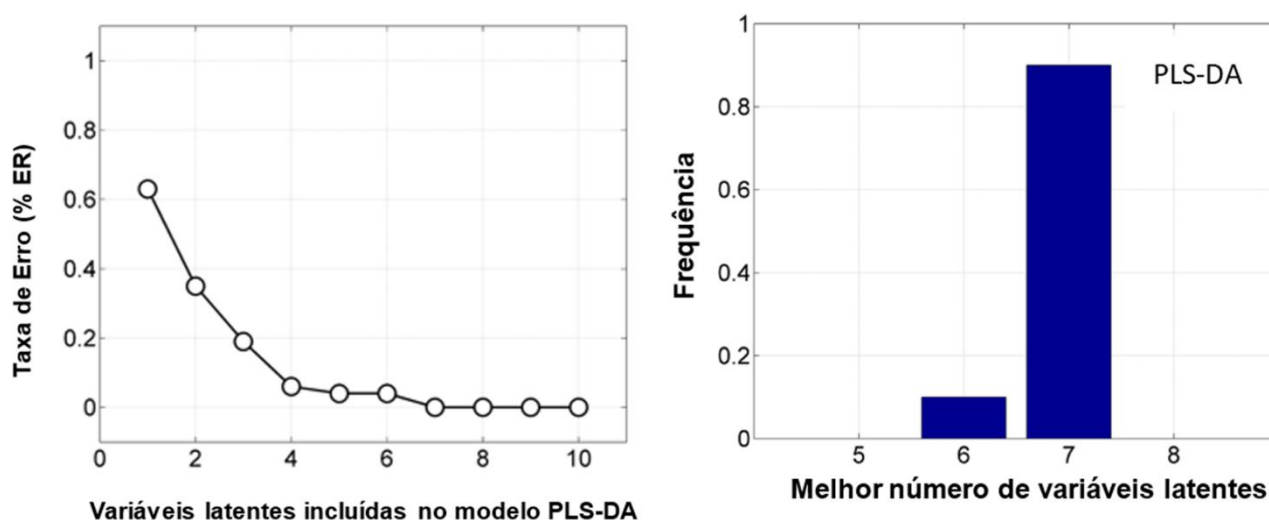


Figura 4.11. Número de variáveis latentes selecionados pelo PLS-DA para classificação de óleos comestíveis.

O desempenho do modelo PLS-DA obtidas com as variáveis latentes podem ser visualizadas através dos escores mostrados na **Figura 4.12 (A)**, bem como o gráfico dos pesos (**Figura 4.12(B)**) das variáveis latentes escolhidas.

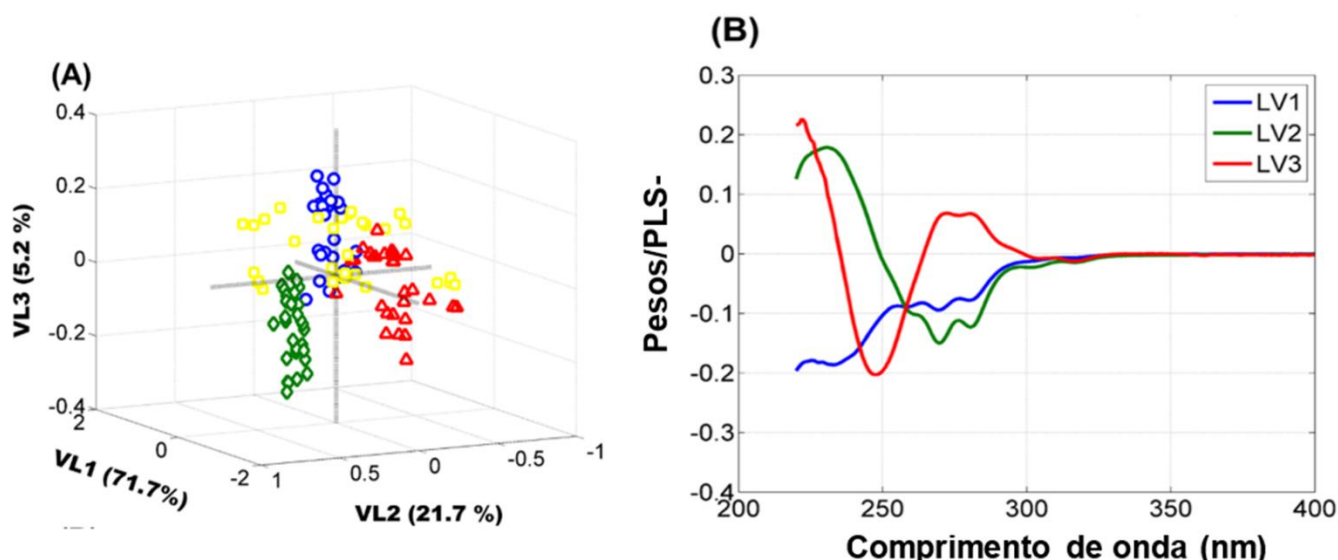


Figura 4.12: Gráfico de escores (A) e pesos (B) do PLS-DA. Círculo azul (canola), triângulo vermelho (soja), quadrado amarelo (milho) e diamante verde (girassol).

No entanto, uma separação de classes clara não é mostrada no gráfico de pontuação do modelo PLS-DA. Vale ressaltar que os modelos LDA construídos no domínio espectral (usando as variáveis selecionadas) fornecem uma conexão direta entre as direções de variância e discriminação. Por outro lado, a interpretação dos carregamentos dos modelos PLS-DA é uma tarefa mais complexa.

Diante dos diferentes modelos de classificação para os diferentes tipos de oleaginosas dos óleos comestíveis, são ressaltados na **Tabela 4.1** os resultados obtidos pelos modelos ACO-LDA, GA-LDA e PLS-DA para o conjunto de treinamento e validação das amostras de óleos. Como pode ver, todos os modelos foram satisfatórios em termos de ajuste e validação. Modelos altamente seletivos e específicos foram obtidos. Apenas o modelo PLS-DA fez uma classificação errada na etapa de Treinamento, onde uma amostra da classe 1 (milho) foi designada como membro da classe 4 (girassol). Portanto, os modelos baseados na seleção de variáveis mostraram uma precisão ligeiramente superior ao modelo completo PLS-DA baseado na etapa de validação. Uma comparação entre ACO e GA mostra resultados semelhantes em termos de precisão. Note que o ACO selecionou apenas 7 variáveis, resultado concordante com o PLS-DA que

apresentou alto desempenho com 7 variáveis latentes. No entanto, por outro lado, o GA parece ter selecionado um número excessivo de variáveis.

Tabela 4.1. Parâmetros de adequação e validação dos modelos PLS-DA, GA-LDA e ACO-LDA empregados para a classificação de óleos vegetais comestíveis.

	Parâmetros					
	Treinamento			Validação		
Modelos	ACC (%)	SEN (can-soj-mil-gir)	SPEC (can-soj-mil-gir)	ACC (%)	SEN (can-soj-mil-gir)	SPEC (can-soj-mil-gir)
PLS-DA (7)*	98	1-1-0.95-1	1-1-1-0.95	100	1-1-1-1	1-1-1-1
GA-LDA (17)*	100	1-1-1-1	1-1-1-1	100	1-1-1-1	1-1-1-1
ACO-LDA (7)*	100	1-1-1-1	1-1-1-1	100	1-1-1-1	1-1-1-1

Sensibilidade (SEN); especificidade (SPEC); precisão (ACC); can (canola); soj (soja); mil (milho); gir (girassol). * O número ideal de variáveis latentes ou selecionadas é indicado entre parênteses.

Através dos resultados da **Tabela 4.1**, é possível determinar a sensibilidade e a especificidade do modelo construído. A sensibilidade do modelo corresponde à sua habilidade em prever corretamente todas as amostras de validação de uma classe em particular. Se todas as amostras de validação de uma determinada classe forem classificadas corretamente, então a sensibilidade do modelo para aquela classe é igual a 1. No modelo construído pelo PLS-DA a sensibilidade para a classe de girassol foi igual a 1 enquanto que para a classe milho foi igual a 0.95. Isso significa que o modelo não foi capaz de classificar corretamente todas as amostras da classe milho. A especificidade corresponde à previsão incorreta de amostras de validação de outras classes em uma determinada classe. Dessa forma, se o modelo não apresentar erros de previsão a especificidade do modelo será igual a 1. Na classe canola a especificidade foi igual a 0.94, pois uma amostra de validação da classe soja foi prevista na classe canola. Por outro lado, a especificidade para a classe milho foi igual a 1 devido a nenhuma amostra de validação da classe soja ter sido prevista como sendo pertencente à classe girassol.

Portanto, os modelos baseados na seleção de variáveis mostraram uma precisão ligeiramente superior ao modelo completo PLS-DA baseado na etapa de treinamento.

Para avaliar o poder preditivo, os modelos validados foram usados para atribuir as amostras de óleo vegetal comestível a um conjunto de teste independente. A **Tabela 4.2** apresenta o resumo dos erros de classificação para o conjunto das amostras externas das três estratégias de modelagem empregadas nesse estudo.

Tabela 4.2: Matriz confusão de predição dos óleos comestíveis

	ACO-LDA				GA-LDA				PLS-DA			
	can	soj	mil	gir	can	soj	mil	gir	can	soj	mil	gir
can (11)	11	-	-	-	11	-	-	-	11	-	-	-
soj (12)	-	12	-	-	-	10	2	-	-	12	-	-
mil (11)	-	-	11	-	-	-	11	-	-	-	11	-
gir (13)	-	-	-	13	-	-	-	13	-	-	-	13
SEN	1	1	1	1	1	0.83	1	1	1	1	1	1
SPEC	1	1	1	1	1	1	0.94	1	1	1	1	1
ACC (%)	100				95.7				100			

Sensibilidade (SEN), especificidade (SPEC) e acurácia (ACC) da classe can (canola), classe soy (soja), classe mil (milho) e classe gir (girassol).

Como pode ser visto na **Tabela 4.2**, precisões de 100, 95,7 e 100% foram obtidas pelos modelos ACO-LDA, GA-LDA e PLS-DA, respectivamente. Isso demonstra que o ACO-LDA foi capaz de alcançar resultados tão bons quanto o PLS-DA e consideravelmente melhores que o GA-LDA, sendo que ambos são ferramentas quimiométricas poderosas e versáteis.

4.3.2. Classificação simultânea de chá com respeito a variedades e origens geográficas

4.3.2.1. Perfil espectral das amostras de chás

Na **Figura 4.13** é apresentado o espectro NIR brutos das 100 amostras de chás na faixa de 1001 a 2500 nm.

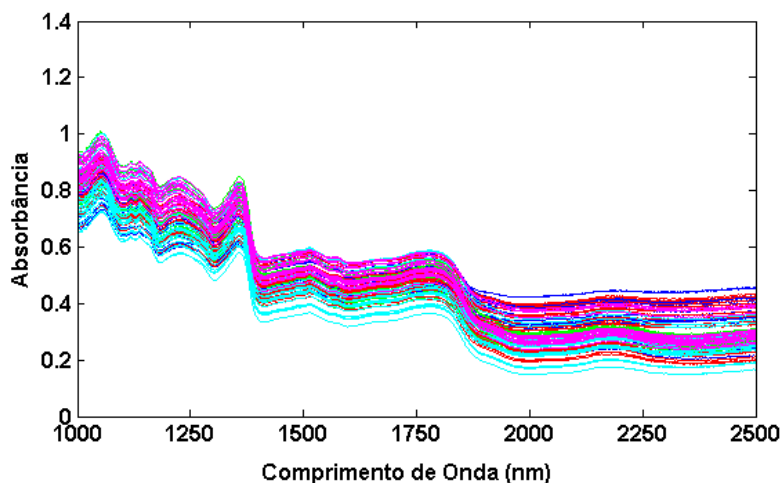


Figura 4.13. Espectros brutos NIR de todas as 5 categorias das amostras de chás analisados.

Na **Figura 4.13** pode-se observar que existem o segundo sobretons de combinações N-H e O-H em 1055 nm e sobretons de estiramento C-H na faixa de 1135-1185 nm. As bandas de absorção de água em torno de 1310 e 1840 nm referente as vibrações de deformação e estiramento de O-H. A banda mais intensa no espectro pertence a vibração de estiramento e sobretom do grupo carbonila (1360 nm), seguindo pelas vibrações de estiramento e deformação C-H (1417nm) e sobretons de CH₂ (1750 nm) e CH₃ (1780 nm). As vibrações do grupo carbonila, CH e CH₂ são causadas por alguns ingredientes, tais como: polifenóis, alcaloides, proteínas, ácidos voláteis e não voláteis e alguns compostos de aroma.

Para analisar melhor o comportamento espectral de cada amostra de chás, é preciso verificar o perfil das classes envolvidas (**Figura 4.14**).

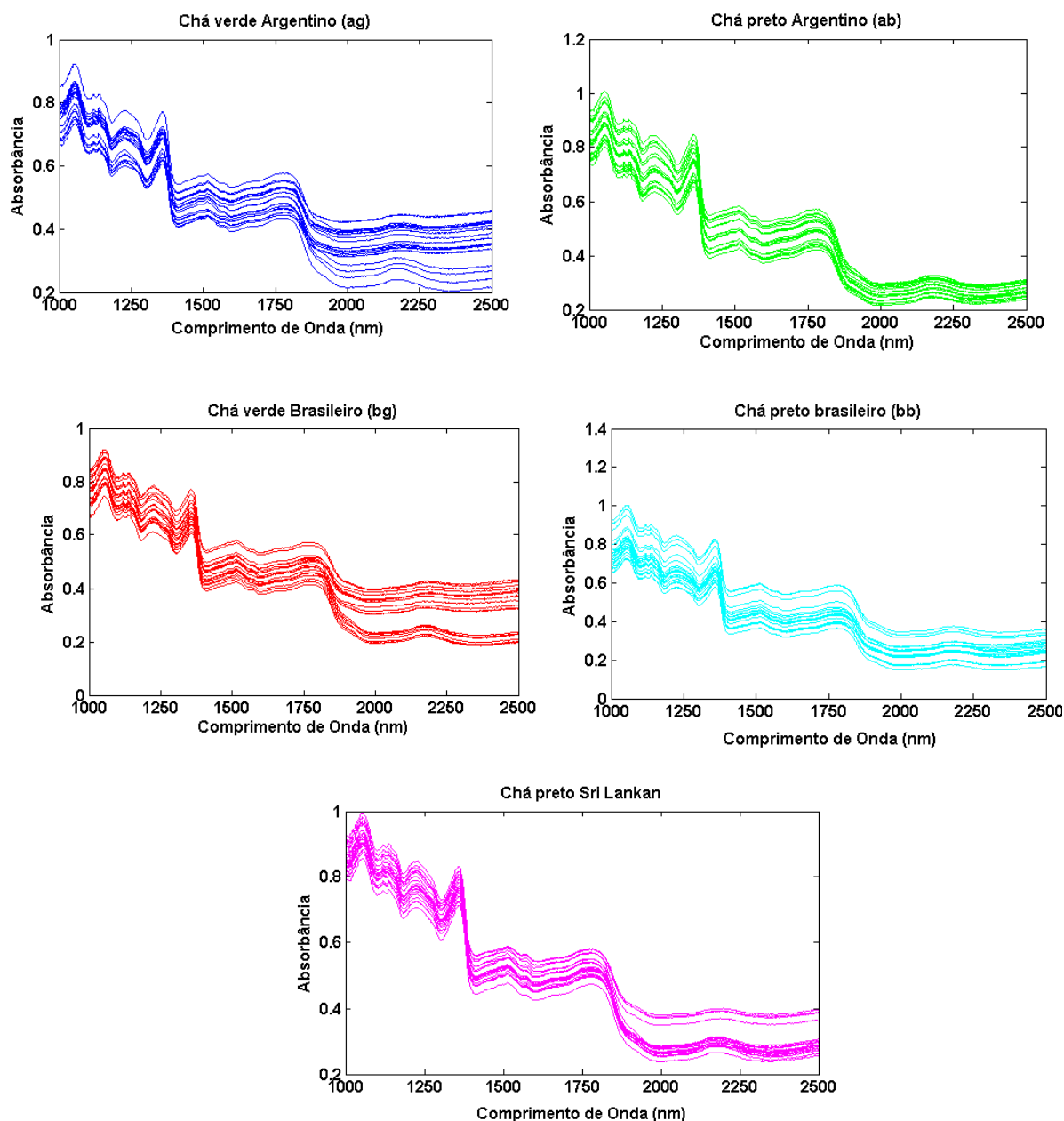


Figura 4.14. Espectros NIR das 5 classes das amostras de chás analisados.

Como pode-se observar, o comportamento das cinco classes são praticamente iguais. Pode-se observar que um dos cinco tipos de chás apresenta um pico de absorção superior de 0.9 ($Abs > 0.9$) enquanto os demais abaixo ($Abs < 0.9$), demonstrando que o principal composto, grupos fenólicos, é mais significativo no chá preto Sri Lanka como observado no gráfico médio das classes (**Figura 4.15**).

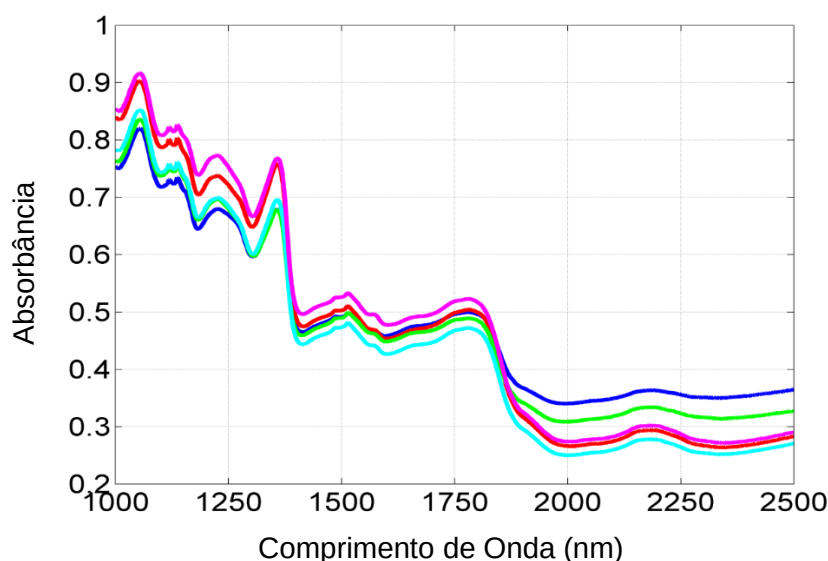


Figura 4.15: Espectros médios de chá de AG (linha azul), AB (linha verde), BG (linha vermelha), BB (linha ciano) e SB (linha magenta)

Para avaliar a capacidade discriminante dos espectros NIR em relação ao tipo e à origem geográfica das amostras de chás, técnicas de classificação foram aplicadas aos espectros brutos das amostras e os resultados são apresentados na seção seguinte.

4.3.2.2. Classificação

4.3.2.2.1 ACO-LDA

Neste estudo de caso, a classificação simultânea do chá em relação ao tipo (verde / preto) e origem geográfica (Argentina / Brasil / Ásia), pelo modelo ACO-LDA foi apresentado como solução ótima, com um subconjunto contendo quinze variáveis espectrais (1398, 1503, 1533, 1570, 1645, 1678, 1790, 1815, 1848, 1860, 2136, 2212, 2329, 2442, 2464 nm) levando em consideração o limite de 0,35 como observado na **Figura 4.16**. Onde pode-se perceber que a inspeção das frequências de seleção nos ciclos de Monte Carlo, permite inferir que o processo de cooperação entre formigas artificiais (mecanismo ausente em AG), leva a uma convergência mais eficiente.

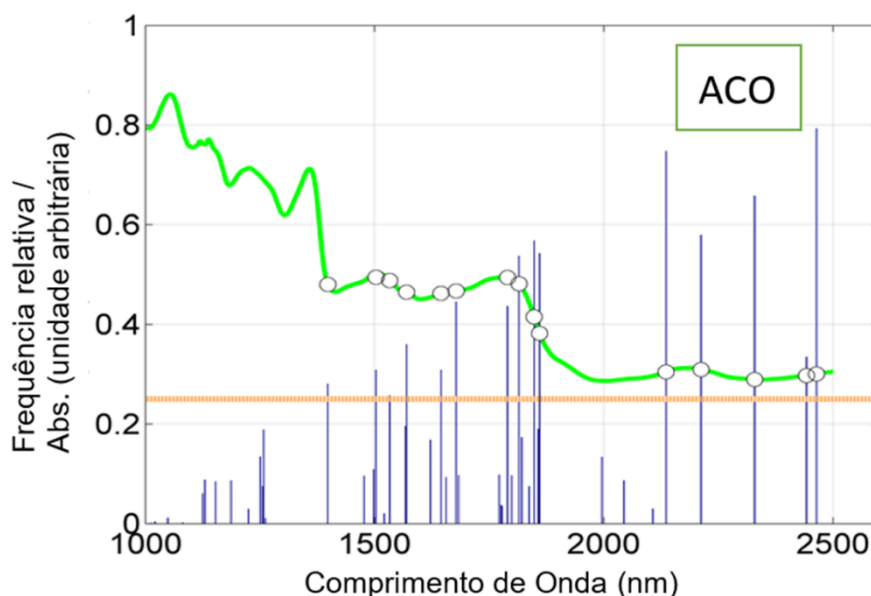


Figura 4.16: Comprimento de onda ou variáveis latentes (círculos brancos) selecionados pelos algoritmos ACO-LDA nos espectros NIR.

Na **Figura 4.16** pode-se notar que dos quinze comprimentos de onda selecionados pelo ACO-LDA, doze estão associados a pontos característicos dos espectros das 5 categorias de chás que portam informações suficientes (bandas de combinações, primeiro e segundo sobretons) para distingui-las e três em regiões de baixa intensidade de absorção.

Da mesma forma do conjunto de dados anterior (óleos comestíveis), a **Figura 4.17** representa os valores de VIF para variáveis preditivas selecionados por ACO-LDA para o conjunto de dados de chás. O uso da análise de inflação de variância foi utilizada como critério para minimizar a multicolineariedade entre as variáveis menos correlacionadas. Esse processo assegura que as variáveis finais selecionadas tenham multicolinearidade desprezível. E pode-se observar que exibe valores de VIF baixos.

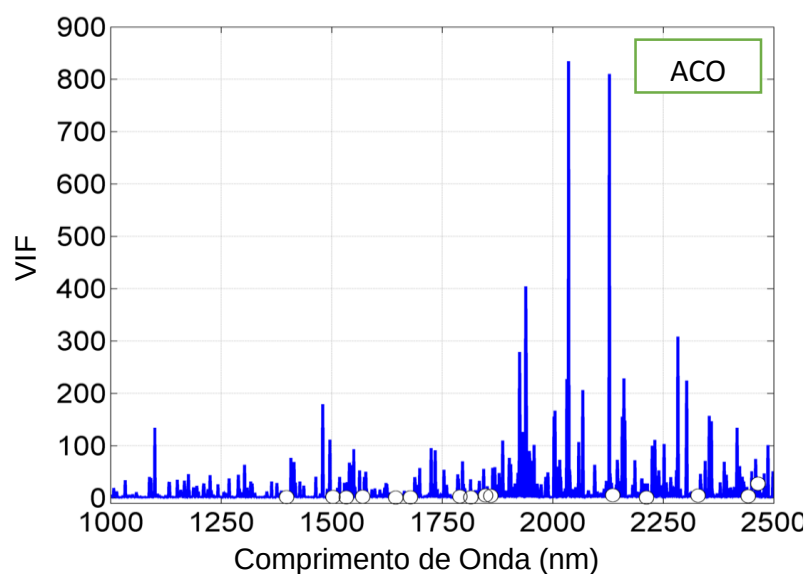


Figura 4.17: Valores VIF (linha azul sólida) para as variáveis selecionadas (círculos brancos com borda preta) por ACO-LDA dos dados do Nir.

A **Figura 4.18** apresenta o gráfico dos escores obtidos pela LDA com prévia seleção de variáveis pelo ACO. É possível observar no gráfico dos escores obtido pelo LDA/ACO uma discriminação entre as classes de chás quanto à sua variedade e origem geográfica.

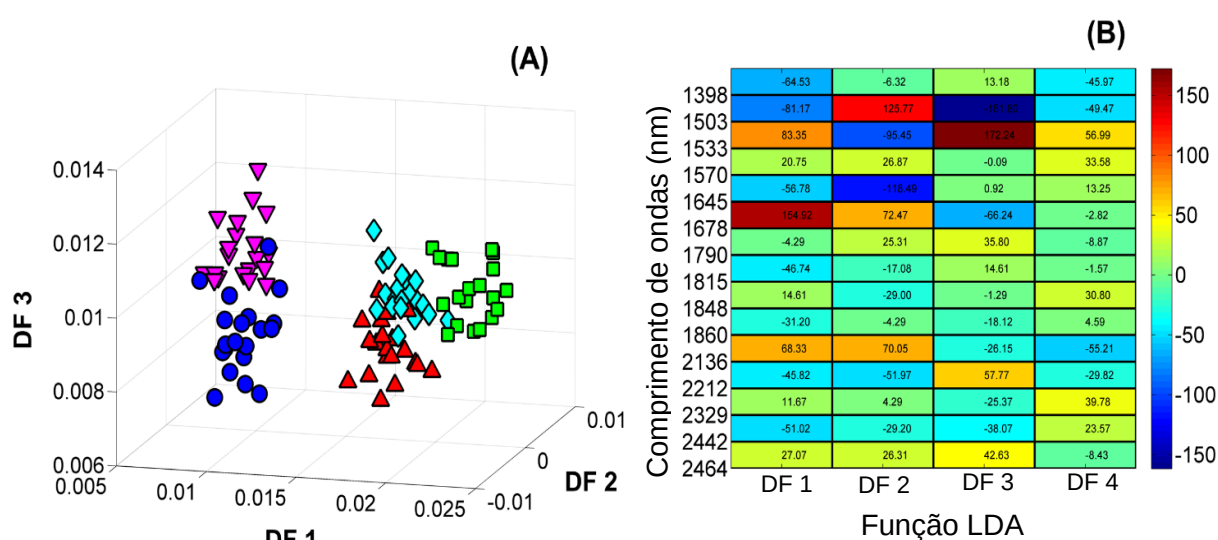


Figura 4.18: Gráfico de escores e mapa de calor de pesos dos modelos ACO-LDA (A e B), para classificação de chás. AG (círculo azul), AB (quadrado verde), BG (triângulo vermelho), BB (diamante ciano) e SB (triângulo margenta de cabeça para baixo).

As amostras de chá verde argentino e verde brasileiro, por sua vez, encontram-se discriminadas das demais classes dos chás pretos em FD2. Nota-se que há uma separação entre as cinco classes analisadas, que pode ser atribuída à capacidade do modelo ACO-LDA de selecionar variáveis com mínima colinearidade ou seja os comprimentos de onda que mais contribuíram para a discriminação e classificação dos chás. Neste gráfico observa-se a formação dos cinco grupos de amostras correspondentes as duas categorias de chás (verde e preto) e as três classes correspondendo a origem geográfica (brasileiro, argentino e srilanquense). Isso está relacionado com o peso que cada comprimento de onda tem para que ocorra a distinção das classes. Diante disso, na **Figura 4.18 B**, observa-se que entre as variáveis escolhidas pelo modelo ACO-LDA, as variáveis de maiores pesos para FD1 foram os comprimentos de ondas de 1503, 1533, 1678 e 2136 nm, para o FD2 foram 1503, 1533, 1645, 1678 e 2136 nm e para o FD3 foram 1503, 1533 e 1678 nm. Logo, as funções discriminante que obtiveram as melhores discriminação das classes foram FD1 e FD2. Tais comportamentos é observado devido a composição química dos chás variar quanto as espécies, idade das folhas, estação, clima e condições de cultivar, bem como do tipo de processamento das folhas dos chás^[133].

4.3.2.2.2 GA-LDA

Para propósito de comparação, o GA foi também empregado para seleção de variáveis em LDA para verificar o comportamento das variáveis selecionadas pelo algoritmo. A **Figura 4.19** mostra os espectros médios das classes com a indicação dos vinte e seis comprimentos de onda selecionados pelo GA-LDA mesmo com o valor limite de 0,5. As variáveis selecionadas correspondem às medidas espectroscópicas nas absorbâncias de 1692, 1795, 1832, 1858, 1968, 2040, 2052, 2053, 2077, 2089, 2111, 2137, 2250, 2310, 2311, 2334, 2335, 2342, 2384, 2424, 2425, 2447, 2490, 2491, 2495, 2500 nm. Como é observado, o GA selecionou uma quantidade grande de variáveis comparadas ao ACO.

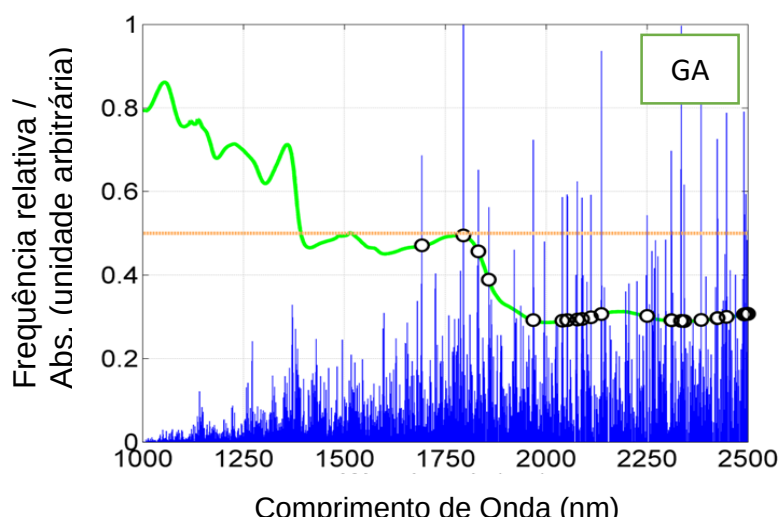


Figura 4.19: Comprimento de onda ou variáveis latentes (círculos brancos) selecionados pelos algoritmos GA-LDA nos espectros NIR.

Observe que para GA as variáveis não informativas têm uma alta frequência de seleção, conforme claramente observado na Figura. Pode-se inferir que o GA selecionou na faixa acima de 2300 nm do espectro de NIR bruto, onde têm informações químicas pobres e/ou apresentam apenas ruídos.

O valor VIF para variáveis preditivas selecionadas por GA-LDA (**Figura 4.20 A**) exibe valores VIF um pouco mais altos para aquelas selecionadas por ACO-LDA (**Figura 4.20 B**), sugerindo que GA-LDA pode estar selecionando variáveis redundantes.

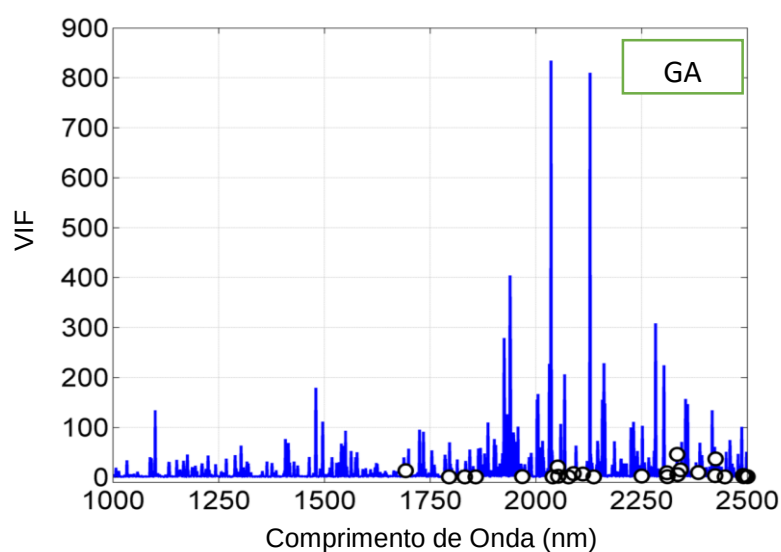


Figura 4.20: Valores VIF (linha azul sólida) para as variáveis selecionadas (círculos brancos com borda preta) por GA-LDA dos dados do Nir.

O modelo GA-LDA resultou na seleção de vinte e seis comprimentos de onda. E comparando com o número de variáveis selecionadas no modelo ACO-LDA, o GA-LDA apresenta-se menos parcimonioso e resultou em um tendência de separação menos efetiva, ou seja, há uma maior dispersão das amostras, o que leva à uma sobreposição entre as classes como observado na **Figura 4.21 A**.

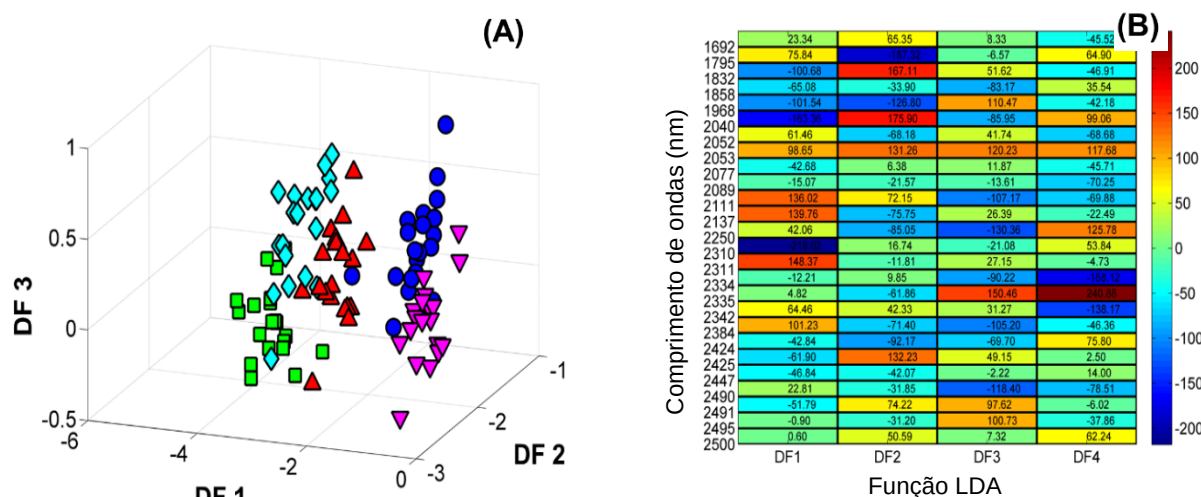


Figura 4.21: Gráfico de escores e mapa de calor de pesos dos modelos GA-LDA (A e B), para classificação de chás. AG (**círculo azul**), AB (**quadrada verde**), BG (**triângulo vermelho**), BB (**diamante ciano**) e SB (**triângulo magenta de cabeça para baixo**).

A partir o gráfico dos pesos na **Figura 4.21 B**, observa-se que entre as variáveis escolhidas pelo modelo GA-LDA, as variáveis de maiores pesos para FD1 foram as variáveis 2040 e 2310 nm e para FD2 foram 1795, 1832, 2040 nm. Essas variáveis foram as melhores que contribuíram para o comportamento das amostras observadas no gráfico dos escores.

4.3.2.2.3 PLS-DA

O modelo PLS-DA construído para as cinco classes de chás estudadas foi com o número das variáveis latentes ideais que exibiram a maior frequência de onze fatores como mostrado na **Figura 4.22**.

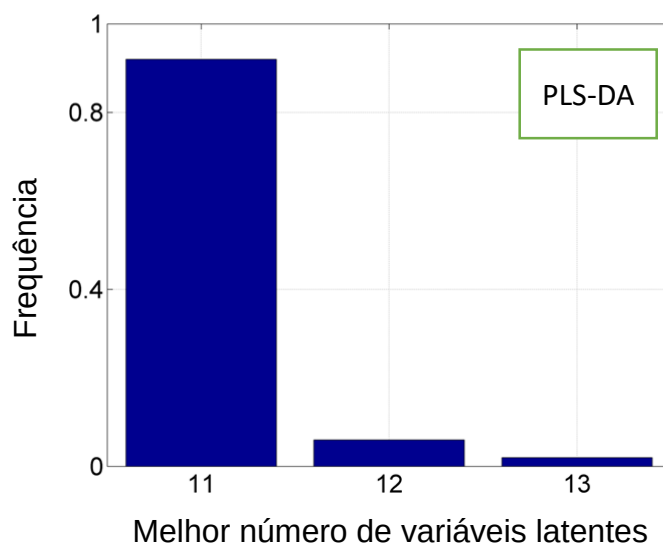


Figura 4.22. Número de variáveis latentes selecionados pelo PLS-DA para classificação de chás.

Os dois tipos de chás estudados são diferenciados de acordo com o grau de fermentação e oxidação das folhas, o que pode corroborar para a separação das classes observadas no gráfico dos scores na **Figura 4.23 A**. Neste caso, observa-se que, embora a terceira variável latente explique apenas 3,5% essa componente é importante para a separação observada.

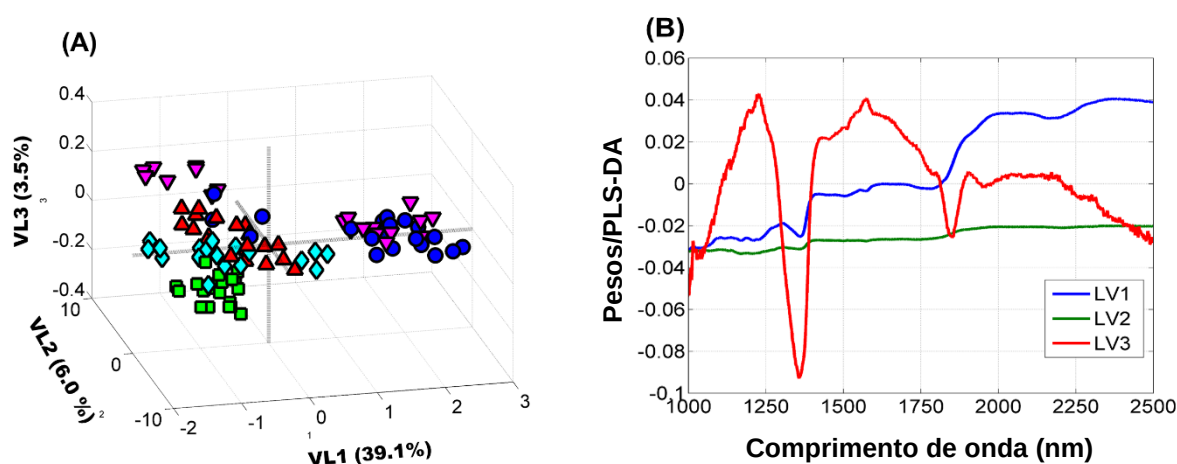


Figura 4.23: Gráfico de escores (A) e pesos (B) dos modelos ACO-LDA (A e B), para classificação de chás. AG (**círculo azul**), AB (**quadrada verde**), BG (**triângulo vermelho**), BB (**diamante ciano**) e SB (**triângulo margenta de cabeça para baixo**).

A análise espectral realizada com base nos gráficos dos pesos (loadings) para as três variáveis latentes (**Figura 4.23 B**) indicam que a separação das classes de **AB** e **AG** se deve, principalmente, pela VL3. Onde é mostrado a contribuição dos maiores pesos, tanto na região positiva como negativa referente ao segundo sobreton das ligações C-H.

Os resultados de todos os modelos apresentados na classificação simultânea de chá com respeito a variedades e origens geográficas são mostrados na **Tabela 4.3** de forma resumida a qualidade de ajuste e validação de cada modelo. Ajuste satisfatório e precisão foram obtidos para todos os modelos, indicando que os espectros NIR carregam as informações químicas que permitem a discriminação entre os chás com relação à variedade e origem geográfica. Pode verificar que no conjunto de treinamento os modelos GA-LDA e ACO-LDA obteve taxas de precisão maiores do que no PLS-DA e no conjunto de validação, a maior taxa de precisão foram os do modelo ACO-LDA.

Tabela 4.3: Parâmetros de adaptação e validação dos modelos PLS-DA, GA-LDA e ACO-LDA para classificação de chás

	Parametros					
	Treinamento			Validação		
Modelos	TCC (%)	SEN (ag-ab-bg-bb-sb)	SPEC (ag-ab-bg-bb-sb)	TCC (%)	SEN (ag-ab-bg-bb-sb)	SPEC (ag-ab-bg-bb-sb)
PLS-DA (11)	97	0.93-1-1-1-0.93	1-1-0.98-0.98-1	91	0.95-0.95-0.95-0.95-0.76	0.95-0.95-0.95-0.90-0.95
GA-LDA(26)	100	1-1-1-1-1	1-1-1-1-1	92	1-1-1-0.86-0.6	1-1-1-0.95-0.95
ACO-LDA (15)	100	1-1-1-1-1	1-1-1-1-1	99	1-1-1-0.92-1	1-1-1-1-0.91

Sensibilidade (SEN), especificidade (SPEC) e precisão (ACC) da classe ag (chás verdes argentinos) - classe ab (chás pretos argentinos) - classe bg (chás verdes brasileiros) - classe bb (chás pretos brasileiros) - classe sb (Chás pretos do Sri Lanka). * Os valores de mérito da validação são a média de todos os ciclos de Monte Carlo.

Vale ressaltar que o modelo ACO-LDA atingiu resultados semelhantes aos obtidos pelo GA-LDA na etapa de treinamento. Por outro lado, na validação por ciclos repetidos de MC, mostrou-se superior. Além

disso, com significativamente menos variáveis, ou seja, o modelo ACO-LDA é mais parcimonioso que o GA-LDA. Além disso, é importante mencionar que de forma geral os resultados do método proposto (para as etapas de treinamento e validação) foram ligeiramente melhores que o modelo PLS-DA.

Os modelos ACO-LDA, GA-LDA e PLS-DA validados foram então aplicados para classificar amostras de chás em um conjunto de teste independente; os resultados são mostrados na **Tabela 4.4**. PLS-DA mostrou os melhores resultados gerando respectivos valores de sensibilidade (SEN), especificidade (SPEC) e precisão (ACC) de 1, 1 e 100%.

Tabela 4.4: Matriz de confusão para previsão de chás em um conjunto de teste independente.

	ACO-LDA					GA-LDA					PLS-DA				
	ag	ab	bg	bb	sb	ag	ab	bg	bb	sb	ag	ab	bg	bb	sb
ag (5)	5	-	-	-	-	4	-	-	1	-	5	-	-	-	-
ab (5)	-	5	-	-	-	-	4	-	1	-	-	5	-	-	-
bg (5)	-	-	5	-	-	-	-	5	-	-	-	-	5	-	-
bb (5)	-	1	-	4	-	-	-	-	4	1	-	-	-	5	-
Sb (5)	-	-	-	1	4	-	-	-	1	4	-	-	-	-	5
SEN	1	1	1	0.8	0.8	0.8	0.8	1	0.8	0.8	1	1	1	1	1
SPEC	1	0.95	1	0.95	1	1	1	1	0.85	0.95	1	1	1	1	1
ACC (%)	92					84					100				

Sensibilidade (SEN), especificidade (SPEC) e precisão (ACC) da classe ag (chás verdes argentinos) - classe ab (chás pretos argentinos) - classe bg (chás verdes brasileiros) - classe bb (chás pretos brasileiros) - classe sb (Chás pretos do Sri Lanka).

Alcançando resultados semelhantes ao PLS-DA, bem estabelecido, o método proposto exibiu um excelente resultado, considerando o pequeno subconjunto de variáveis selecionadas (1% de todas as informações espectrais). A precisão na predição foi de 92% para ACO-LDA enquanto para GA- LDA foi de apenas 84%.

CAPÍTULO V

CLASSIFICAÇÃO DE TRÊS CULTIVARES DE FEIJÃO



5. CLASSIFICAÇÃO DE TRÊS CULTIVARES DE FEIJÃO

5.1. Introdução

5.1.1 Feijão

O feijão é uma planta herbácea, pertencente à família *Fabaceae*, que compreende as leguminosas, sub-família *Faboideae*, gênero *Phaseolus* e espécie *Phaseolus vulgaris* L.. É uma leguminosa comestível muito estudada em todo o mundo, por contribuir com 95% da produção mundial de feijões¹⁴². Conforme dados registrados pela FAO (Food and Agriculture Organization) a produção média mundial de feijão no período de 2010 a 2014 foi de 23,9 milhões de toneladas ¹⁴³. Índia e Mianmar (16%), Brasil (13%), EUA, México e Tazania (5%) e China (4%) foram responsáveis por 64% do total produzido no período¹⁴³.

No Brasil, o feijão (*Phaseolus vulgaris* L.) ocupa lugar de destaque na agricultura brasileira, sendo caracterizado como forte produto no mercado interno, cujos grãos representam uma importante fonte de proteína e minerais na dieta diária da população¹⁴⁴. Estados do Paraná, Minas Gerais, Bahia, São Paulo, Goiás, Ceará, Pernambuco, Rio Grande do Sul, Santa Catarina e Pará são os 10 principais produtores dessa leguminosa responsáveis por praticamente 85% da produção nacional, atingindo anualmente cerca de 3,0 milhões de toneladas, distribuídas em três safras distintas, águas, seca e inverno ¹⁴³.

Recentemente a produtividade média da cultura, na safra de 2012/2013 foi de 3.437 milhão de toneladas considerando as três safras anuais. Porém, a média de produtividade da primeira safra teve médias de 1.255 milhão de toneladas, na segunda safra 1.315 milhão de toneladas e na terceira safra 866 milhão de toneladas ¹⁴⁵. Atualmente, a previsão da companhia nacional de abastecimento (Conab) é de que os produtores aumentem a área plantada de feijão, passando de 2,8 milhões de hectares na safra 2015/16 para 3 milhões de hectares em 2017/2018.¹⁴⁶

Os dois gêneros de feijões cultivados no Brasil são *Phaseolus* e o *Vigna*, sendo que o primeiro é mais cultivado na região Centro Sul (carioca e preto), e o segundo na região Norte/Nordeste (macaçar/caupi)¹⁴⁷. As variedades mais produzidas são o feijão Cariquinha e em seguida pelo feijão preto e em pequenas quantidades “outras variedades” que são os feijões, vermelho, canário, jalo, rajado e rosinha, atendendo alguns nichos no mercado interno e externo ¹⁴⁷.

A exigência do mercado quanto à cor e ao tipo de grãos é variável de região para região. Por exemplo, no Rio Grande do Sul, em certas regiões de Santa Catarina e Paraná, Rio de Janeiro e Espírito Santo, a preferência é pelo feijão preto. Em São Paulo e em algumas regiões de Minas Gerais a preferência é pelo feijão de cor, principalmente o mulatinho, roxinho e pardo ¹⁴⁸. No estado de Goiás são cultivados vários tipos de feijões, mas predomina o tipo carioca. Em Minas Gerais são recomendados alguns cultivares como perola, carioca MG dentre outros ¹⁴⁹. Na faixa que vai da Bahia até o Rio Grande do Norte, é onde está concentrada a produção da variedade *Vigna*, sendo que na região Norte (RO), a maior parte da área é destinada ao plantio de carioca, com destaque para o estado do Pará, concentrando uma boa produção do feijão caupi e mais recentemente TO, produzindo o carioca, com custos mais baixos, atraindo produtores de GO e MG ¹⁴³.

A Embrapa conta com ampla variedade de cultivares desenvolvidas para as diferentes condições de clima e solo do País, são elas: BRS Estilo, BRS Requite, Perola, BRS Notável, BRS Amestita, BRS Agreste, BRS Estilo, BRS pontal, BRSFC401RMD, BRSMG Realce, BRS Esteio, BRS Campeiro, BRS Pitanga, RBS União, BRS Requite e várias outras variedades. Dentre essas variedades, serão aplicadas no trabalho as cultivares BRS Estilo, BRS Pontal e BRSFC401RMD. Este último é o cultivar do feijão transgênico. O perfil de cada um será apresentado a seguir:

5.1.1.1. Cultivar BRS Estilo¹⁵⁰

A cultivar BRS Estilo diferencia-se pela arquitetura de plantas ereta, de crescimento indeterminado tipo II, adaptada à colheita mecânica direta, possui ciclo normal de 85 a 95 dias e apresenta massa de 100 grãos em torno de 26g. Apresenta alto potencial produtivo e estabilidade de produção, com grãos claros e de excelente qualidade comercial. Em relação às doenças, a cultivar BRS Estilo apresenta resistência ao mosaico comum e reação intermediária à antracnose e ferrugem.

Os feijões com crescimento indeterminado, caracteriza-se por possuir um caule principal com tecido embrionário existente na planta na fase vegetativa (meristema) que permite um crescimento contínuo, numa sucessão de nós e entrenós; as inflorescências desenvolvem-se nas axilas das folhas, isto é, a partir do ponto de inserção de uma folha ou ramo na haste principal ou secundária de uma planta (inflorescência axilar) e a floração inicia-se da base para o ápice da planta

5.1.1.2. Cultivar BRS Pontal¹⁵¹

A BRS Pontal é uma cultivar desenvolvida pela Embrapa, de alto potencial produtivo e arquitetura de planta semi prostrado. E apresenta massa de 100 grãos ao redor de 26 g e ciclo normal. A BRS Pontal é uma cultivar com tipo de grão comercial carioca, ciclo normal (85 a 95 dias). Esta cultivar de feijoeiro comum possui, ainda, como características boa resistência a doenças, principalmente à antracnose (11 raças), tendo reação intermediária a ferrugem e crestamento bacteriano comum e é suscetível à mancha angular. Tem como público-alvo agricultores e agricultura familiar, as cooperativas, associações de produtores rurais e empresas agropecuárias.

5.1.1.3. Cultivar BRSFC401RMD¹⁵²

A Cultivar BRS FC401 RMD apresenta ciclo normal (média de 87 dias da emergência à maturação fisiológica, variando de 85 a 91 dias), bom potencial produtivo, plantas com arquitetura prostrada, com hábito de crescimento indeterminado (tipo III) e os grãos são bege-claros com estrias marron-claras de forma elíptica semicheia e sem brilho. A BRS FC401 RMD não apresenta tolerância ao acamamento, sendo recomendada, preferencialmente, a colheita mecânica em duas etapas (corte e trilha separados). As flores são brancas e, na maturação fisiológica, as vagens são amarelo-avermelhadas. Já na maturação de colheita, as vagens têm coloração amarela. Possui resistência efetiva ao mosaicodourado, por meio da transgenia (tecnologia RMD), resistência ao mosaico-comum, resistência moderada à antracnose e à ferrugem.

5.1.2. Composição Química

O feijão representa mais de 50% do consumo de grãos na alimentação humana, por ser um excelente fornecedor de nutrientes essenciais como: proteínas, ferro, zinco, cálcio, magnésio, zinco, vitaminas (principalmente do complexo B), carboidratos e fibras^{153,154}. Além de ser uma fonte de nutrientes essenciais, o feijão possui constituintes químicos que são considerados componentes com ação funcional, como os flavonoides, antocianidinas, proantocianidinas e isoflavonas. ^{155,156}.

Os grãos possuem de 20 a 35% de proteína, dependendo dos tratos culturais e da cultivar. O teor proteico e a produção são altos, porém, o valor nutritivo da proteína não é satisfatório por apresentar teores baixos de alguns aminoácidos essenciais limitantes. Esses aminoácidos são os sulfurados, a metionina, a cisteína e a cistina, além do triptofano¹⁵⁷. Em feijões também é observada a presença de compostos fenólicos, que se localizam principalmente no tegumento do grão e em variedades coloridas ¹⁵⁸.

No feijão a composição dos carboidratos varia de 60 a 65%, onde o amido é o principal carboidrato armazenado. A porcentagem de lipídios varia entre 0,8 e 1,5%, de acordo com a variedade, localização, clima e tipo de solo ¹⁵⁹. Em termos de fonte de minerais, uma xícara de feijão cozido pode proporcionar 55% do ferro para homens e 29% para mulheres, 20-25% de fósforo, manganês e magnésio, 20% de potássio e 10% de zinco e cálcio¹⁵⁹.

Os programas de melhoramento genético do feijoeiro visam obter variedades que apresentem alta produtividade, aliada a resistência às doenças, com produção de sementes possuindo forma, tamanho, cor e brilho aceitáveis no mercado. Além disso, os grãos de feijão devem possuir textura macia do tegumento, maior teor de proteínas e minerais¹⁶⁰.

5.1.3 Pragas na cultura do feijoeiro

Os estragos causados pelas as pragas na cultura do feijoeiro podem trazer danos desde a semeadura até quando os grãos estão secos nas vagens ou mesmo armazenados. Devido à diversidade de espécies que ocorrem, praticamente todas as estruturas da planta têm se mostrado suscetíveis a essas pragas¹⁶⁰. O feijoeiro é uma planta que apresenta ciclo curto, ou seja, podem ser cultivadas duas a três vezes no mesmo ano agrícola¹⁶⁰. Em decorrência disso, e da variação estacional nas populações de pragas, condições climáticas, cultivares e práticas de cultivo utilizadas, os prejuízos à cultura variam nas diferentes épocas de plantio e a cada ano¹⁶⁰. De maneira geral, as perdas no rendimento causadas pelas pragas e doenças (**Tabela 5.1**) na agricultura mundial têm sido estimadas na faixa de 33 a 86%, dos quais 13% são devidos a insetos ¹⁶¹.

A necessidade de utilização de agrotóxicos no combate às pragas produz um impacto ambiental negativo e um aumento do custo de produção, que pode até mesmo inviabilizar seu cultivo devido às variações de mercado. No caso específico do feijoeiro comum (*Phaseolus vulgaris* L.)

uma das doenças mais agressivas é o mosaico dourado, transmitido pela mosca branca ¹⁶².

Tabela 5.1. Principais doenças que afetam o feijoeiro.

Doença	Patógenos	Folhas
Antracnose	<i>Colletotrichum lindemuthianum</i>	Lesões que ocorrem inicialmente na face inferior da folha
Ferrugem	<i>Uromyces appendiculatus</i>	Lesões de cor esbranquiçada ou amarelada na face inferior das folhas
Mancha angular	<i>Isariopsis griseola</i>	Coloração pardo-acinzentada na face inferior da folha
Mofo branco –	<i>Sclerotinia sclerotiorum</i>	Manchas aquosas
Mosaico Comum	<i>Bean common mosaic virus</i>	Rugosidade, enrolamento das folhas e frequentemente são menores que as folhas saudáveis.
Mosaico Dourado	<i>Bean golden mosaic virus</i>	Iniciam-se nas folhas mais novas com um salpicamento amarelo vivo

Fonte: Rosolem, 1994.

5.1.4 Mosaico dourado

O mosaico dourado é uma doença do feijoeiro causada pelo vírus da espécie *Bean golden mosaic virus* (BGMV), nomenclatura oficial para o vírus do mosaico dourado do feijoeiro (VMDF), e tem como transmissor o inseto conhecido como mosca branca (*Bemisia argentifolii*) ¹⁶³ como ilustrado na **Figura 5.1**.

O BGMV foi descrito pela primeira vez em 1965 em Campinas (SP). Naquela época, a doença foi considerada de importância secundária ¹⁶⁴. Porém, no decorrer da década de 70 a situação se inverteu devido ao considerável aumento populacional do inseto vetor, resultando em uma disseminação alarmante do BGMV nas lavouras de feijoeiro ¹⁶⁴. Após a década de 70, o vírus se disseminou por todas as regiões com clima favorável a ocorrência, alastrando em 100% de doença na cultura levando a perda de 40% a 100% da produção ¹⁶⁴.



Figura 5.1 Agente transmissor da doença do mosaico dourado (mosca branca).

A mosca é vetora do geminivírus (Família *Geminiviridae*), um grupo de vírus cujo genoma é composto por DNA e que é capaz de infectar várias espécies de plantas retirando a seiva elaborada por elas e abrindo portas para fungos oportunistas e transmissores do vírus¹⁶⁵. O nome "mosaico-dourado" tem relação com o amarelecimento intenso, delimitado pela coloração verde, formando um mosaico verde-amarelo conforme **Figura 5.2** ¹⁶⁵. As folhas das plantas infectadas podem apresentar-se deformadas e manchadas, afetando a qualidade das sementes ¹⁶².



Figura 5.2. Cultivar de feijão afetado pelo mosaico dourado.

Fonte: Embrapa Arroz e Feijão

A mosca transmite o vírus com enorme eficiência, podendo levar poucas semanas para que 100% da lavoura estejam infectadas¹⁶⁶. Ao se alimentar de uma planta infectada, o inseto pode adquirir o vírus do mosaico dourado¹⁶⁶. O inseto infectado, ao se alimentar de uma planta sadia, inocula o vírus e algumas toxinas no sistema vascular do feijão¹⁶⁶. Para controlar a ação dessas moscas, o mais comum é a aplicação de inseticida. Entretanto, o uso intenso das mesmas moléculas e, muitas vezes, não associado a outras técnicas de manejo, tem rapidamente reduzido à vida útil dos inseticidas e selecionado genótipos de mosca-branca resistentes à maioria dos produtos disponíveis no mercado, limitando, assim, a eficiência do controle químico num país que chega a plantar três safras (safras das águas, da seca e inverno) anuais de feijão¹⁶⁵. Diante disso, para diminuir a quantidade de produtos químicos nas plantações, a empresa brasileira de pesquisa agropecuária EMBRAPA está pedindo a liberação para plantio de uma variedade de feijão geneticamente modificado capaz de resistir ao vírus do mosaico dourado¹⁶⁶. Esta praga agrícola causa a perda de mais de 90 mil toneladas anuais do produto, podendo chegar a 280 mil toneladas dependendo da época do plantio e da espécie cultivada. Para combater a mosca branca, o vetor de disseminação, os agricultores chegaram a utilizar mais de 20 diferentes tipos de agrotóxicos¹⁷⁰.

5.1.5 Os organismos geneticamente modificados ou transgênicos

Os organismos geneticamente modificados ou transgênicos tende criar organismos com características novas ou melhoradas relativamente ao organismo original, ou seja, o material genético foi alterado de modo artificial: por meio da manipulação genética, combinam-se características de um ou mais organismos de uma forma que provavelmente não aconteceria na natureza¹⁶². OGM são todo e qualquer organismo que teve seu material genético (DNA) modificado por meio de técnicas aplicadas pela engenharia genética, em laboratórios¹⁶⁷. Já os transgênicos são organismos que contêm um ou mais genes transferidos artificialmente de outra espécie¹⁶⁷. Portanto, dentro dos OGM existe o grupo dos transgênicos.

Organismos geneticamente modificados (OGMs), em plantas geneticamente modificadas (PGMs), têm sido desenvolvidos nos últimos anos, a fim de aumentar o fornecimento de alimentos pelo crescimento da produção das colheitas¹⁶⁸⁻¹⁷⁰. Empresas vêm desenvolvendo plantas com características agronômicas melhoradas, onde nos embriões das plantas são inseridos fragmentos de DNA de bactérias, vírus ou fungos que contêm genes que codificam a produção de herbicidas que produzam as toxinas contra as pragas da lavoura, não necessitando de certos agrotóxicos¹⁷¹. Outras são feitas resistentes a certos agrotóxicos, para que estes sejam usados em lavouras onde é preciso exterminar outro tipo de vegetal, como ervas daninhas, sem afetar o resto da produção¹⁷².

Os Estados Unidos, a Argentina, o Canadá e a China são os maiores produtores de transgênicos e correspondem a cerca de 99% da produção mundial¹⁷³. Atualmente, os produtos com organismos geneticamente modificados (OGMs) são cultivados em pelo menos 16 países (Estados Unidos, Argentina, Canadá, China, África do Sul, Austrália, Índia, Romênia, Espanha, Uruguai, México, Bulgária, Indonésia, Colômbia, Honduras e Alemanha), sem contar o Brasil.

Os cultivos de plantas geneticamente modificadas têm sido aprovados no mundo inteiro, dentre os quais podemos destacar o milho, a soja, a canola e o algodão, além do tomate e do mamão, em menor escala^{174,175}.

A cultura com o maior volume de cultivo é a soja, com aproximadamente 60% do total, seguida pelo milho com 21%, pelo algodão com 11% e pela canola com 5% do total cultivado¹⁵⁵. Nos Estados Unidos, mais de 50% do algodão, 45% da soja e 40% do milho cultivados passaram por alguma alteração genética, e pelo menos, 60% dos produtos industrializados contêm OGM ¹⁵⁵.

O Brasil ocupa o segundo lugar entre os países que mais cultivam variedades geneticamente modificadas de grãos e fibras do mundo, perdendo apenas para os Estados Unidos, segundo relatório do Serviço Internacional para aquisição e agrobiotecnologia (ISAAA, na sigla em inglês) ¹⁷⁶. Segundo o Portal do Agronegócio Goiano em 2010, mais da metade de toda a área semeada conta com tecnologia transgênica. Do total de 37,5 milhões de hectares com plantação de soja, milho e algodão, cerca de 67%, ou 25,3 milhões de hectares, correspondem a produção transgênica¹⁷⁶.

Recentemente, em 2016, o Brasil cultivou 49,1 milhões de hectares (ha) com culturas transgênicas em 2016, um crescimento de 11% em relação a 2015 ou o equivalente a 4,9 milhões de ha. Nenhum outro país do mundo apresentou um crescimento tão expressivo. Com essa área, a agricultura brasileira está atrás apenas dos Estados Unidos (72,9 milhões de ha) no ranking global de adoção de biotecnologia agrícola. ¹⁷⁷

No Brasil, pesquisadores da Embrapa Recursos Genéticos e da Embrapa Arroz e Feijão solicitaram à Comissão Técnica Nacional de Biossegurança – CTNBio a liberação para cultivo comercial de variedades de feijão transgênicas, resistentes ao vírus do mosaico dourado. Vale destacar que a pesquisa marca um feito inédito no Brasil, onde as primeiras plantas transgênicas foram totalmente produzidas por uma instituição pública de pesquisa.

5.1.6 O feijão geneticamente modificado

A redução da qualidade comercial e o desempenho de produção dos grãos de feijão estão comprometidos pelo grande número de doenças que

acomete a cultura no país. Entre essas doenças está o mosaico-dourado, principal virose do feijão no Brasil. Seu agente causal é um geminivírus transmitido pela mosca-branca (*Bemisia tabaci*), o Bean golden mosaic vírus (BGMV).

Deste modo, a Embrapa Arroz e Feijão, juntamente com a Embrapa Recursos Genéticos e Biotecnologia, desenvolveram várias linhagens geneticamente modificadas para conter a replicação do vírus BGMV a qual foi identificada como evento Embrapa 5.1¹⁷⁸, a variedade 'Olathe Pinto' portando resistência ao vírus por meio da expressão de RNAs interferentes – RNAi, e comercialmente chamada de tecnologia RMD (resistência ao mosaico-dourado)¹⁷⁸. A resistência da mosca-branca e das viroses por ela transmitidas tem sido a base utilizada pelo Programa de Melhoramento de Feijão da Embrapa para o desenvolvimento de cultivares do grupo carioca com resistência efetiva ao mosaico-dourado, destacando a linhagem BRSFC401RMD¹⁵⁶. Esta cultivar de feijão comum é a primeira registrada e protegida no Brasil com resistência efetiva ao mosaico-dourado, sendo a primeira cultivar geneticamente modificada de feijão-comum já desenvolvida em todo o mundo¹⁵².

A cultivar BRSFC401RMD¹⁵² originou-se do cruzamento entre plantas homozigotas da linhagem de feijão RMD inicialmente desenvolvida, o evento Embrapa 5.1, a qual possuía genótipo da linhagem Olathe Pinto, e plantas da cultivar BRS Pontal, com grãos tipo carioca. Então a linhagens de genótipos dos feijões comerciais BRS Pontal receberam a transgene do cultivar do evento Embrapa 5.1, através de cruzamentos e 4 retrocruzamentos, e também apresentaram completa resistência ao BGMV¹⁵².

Pesquisadores da Embrapa conduziram pesquisas de campo com o feijão transgênico, BRSFC401 RMD, Considerando oito ensaios onde houve ocorrência severa e uniforme de mosaico-dourado (Goiás e Distrito Federal) e apresentaram resistência efetiva ao mosaico-dourado, não apresentando sintomas visíveis, enquanto as testemunhas convencionais, BRS Pontal,

foram suscetíveis, apresentando severidade média de doença de 50% a 70%. ¹⁵²

Na **Figura 5.3** é mostrada uma planta geneticamente modificada, imune ao mosaico dourado (vagens desenvolvidas) e uma planta não modificada, apresentando os sintomas típicos (vagens subdesenvolvidas). Ambas foram inoculadas com o BGMV.

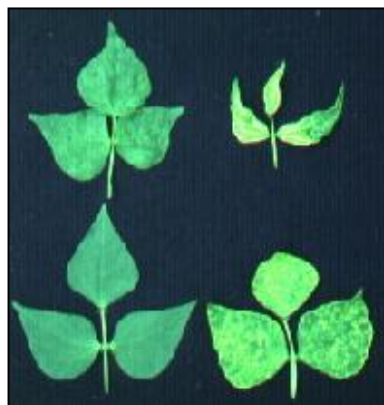


Figura 5.3. Planta geneticamente modificada (esquerda), imune à doença do mosaico dourado e uma planta não modificada (direita).

5.1.7. Ferramentas na detecção dos feijões convencionais e transgênicos

Na detecção de sementes de plantas, prefere-se a utilização de métodos fáceis, tal como o teste de tolerância ou resistência a herbicidas, chamados de bioensaios, que é conduzido através do teste de germinação, de forma similar ao descrito nas Regras para Análise de Sementes do Brasil¹⁷⁹⁻¹⁸⁰. Este teste utiliza testemunhas positivas e negativas, em substrato contendo herbicida específico, sendo que na avaliação final, as características no desenvolvimento, comprimento, coloração, as distingue das provenientes de sementes de plantas GM¹⁷⁹⁻¹⁸⁰.

Para determinação de organismos geneticamente modificados ou transgênicos também são utilizados os métodos convencionais, tais como os baseados em PCR (Protein Chain Reaction - Reação de Polimerização em Cadeia) e ensaios ELISA (enzyme-linked immunosorbent assays - ensaio imunoenzimático)¹⁸⁰⁻¹⁸³. A PCR determina o número de inserções

e de cópias das sequências de DNA introduzidas, além de se pesquisar quais os sítios de inserção dos transgenes e a ELISA detecta proteínas específicas contidas nas sementes de plantas GM¹⁸⁰⁻¹⁸¹.

Outras técnicas analíticas empregadas para determinação organismos geneticamente modificados ou transgênicos podem-se citar: técnica de H¹H¹-MAS RMN¹⁵⁵, eletroforese microchip (ME)¹⁸⁴, quantificada diretamente empregando-se cromatografia líquida de alta eficiência no modo fase reversa (RP-HPLC)¹⁸⁵, dentre outros.

A espectrometria NIR vem se tornando uma metodologia eficiente para as análises de OGM devido às suas vantagens: rapidez de análise, a não necessidade de preparação complexa da amostra ou de processamento e por ser uma técnica não destrutiva. Para especificamente para tentativas de distinguir alimentos transgênicos a partir de alimento convencional, alguns trabalhos podem também ser encontrados na literatura¹⁸⁶⁻¹⁸⁸.

5.2. Experimental

5.2.1. Amostras

Um total de 2700g de amostras de sementes de feijões de diferentes cultivares (900 g do transgênico BRS FC401 RMD, 900g do convencional-genitor: BRS Pontal e BRS Estilo) foram cedidas pela Estação Experimental da Embrapa Arroz e Feijão localizada em Santo Antônio de Goiás-GO, Brasil.

As sementes dos feijões transgênicos foram acondicionadas em embalagem duplo de papelão para melhor conservação, enquanto para sementes dos convencionais e BRS Estilo foi usado embalagem plástica. Em cada pacote, contendo 900 g cada, foram retiradas 30 amostras de cada pacote em porções de aproximadamente 25g e acondicionadas da mesma forma como foram recebidas. De cada porção, foram retiradas 3 sementes aleatoriamente e separadas individualmente totalizando 90 sementes de

cada amostra. A **Figura 5.4** ilustra o processo de separação das sementes para a coleta dos dados.



Figura 5.4. Separação das sementes das diferentes cultivares de feijões

5.2.2 Aquisições dos Espectros reflectância na região do NIR

As medidas de reflectância difusa foram realizadas em um espectrômetro portátil DLP NIRscan Nano EVM (*Evaluation Module*)¹⁸⁹ de aproximadamente 62 mm de comprimento, 58 mm de largura e 36 mm de altura conforme mostrada na **Figura 5.5**. Trata-se de um equipamento capaz de obter informações de absorção referente às vibrações de moléculas orgânicas após interagir com radiação infravermelha na faixa de 900 a 1.700 nm. Como a espectroscopia NIR tem grande aplicação em alimentos, os grupos funcionais presentes nos alimentos como C-H, N-H, e

O-H são detectados através de vibrações eletromagnéticas na região do Infravermelho Próximo ¹⁹⁰



Figura 5.5. espectrômetro portátil DLP NIRscan Nano EVM

O espectrômetro portátil DLP NIRscan Nano EVM foi conectado ao notebook, processador Intel® Core™ i5-2410M de 2.3 GHz e memória RAM de 6 Gb, por intermédio da porta USB. A obtenção dos espectros foram realizados no programa NIRscan Nano GUI v1.1.9 do próprio instrumento como observado na **Figura 5.6.**

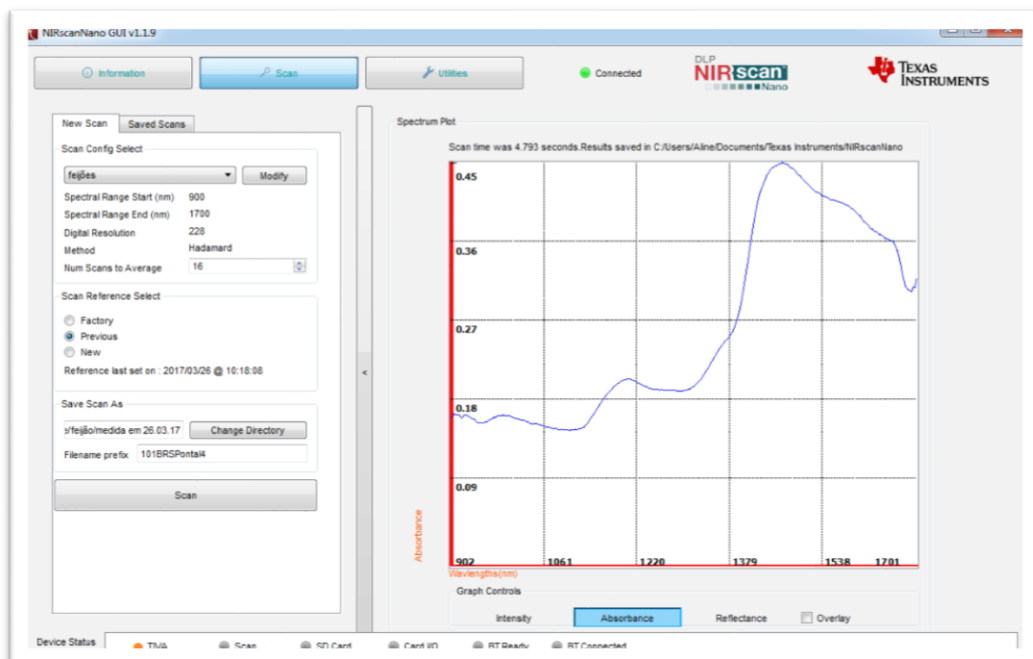


Figura 5.6. Programa NIRscan Nano GUI v1.1.9

O botão scan na parte superior, exibe os gráficos dos espectros e os controles para digitalização, configurações e parâmetros. A faixa de trabalho foi de 900 a 1700nm com resolução de 4nm e foram realizados 16 varredura por amostras utilizando o método hadarmard com largura de 2,00nm e resolução digital de 228.

A **Figura 5.7** mostra uma célula de *Teflon*® no qual foi construída para posicionar as sementes dos feijões a serem analisados em uma posição fixa, por meio de um orifício que permitisse a passagem da radiação NIR. Então, são colocadas as amostras de sementes na célula de medida em uma mesma posição a fim de respeitar a uniformidade das medidas. Essas medidas foram realizadas de forma aleatória.



Figura 5.7. (A) Célula de *Teflon*

Antes da aquisição dos espectros, fez-se o branco utilizando um padrão de reflectância microfibras *Teflon*® em pó, tendo em sua composição polímeros fluorados. O padrão foi adicionado em um béquer de 10ml até que atingisse uma quantidade suficiente para a radiação não passar. (**Figura 5.8**). Essa foi a melhor maneira que encontramos para bloquear a perda da radiação e realizarmos a medida do branco. Pois na reflectância, a referência é um corpo branco que reflete toda a radiação incidente.



Figura 5.8. Padrão de reflectância

Depois de realizada a medida do branco, as aquisições das informações nas amostras foram registradas, em triplicatas, de uma maneira na qual obtivéssemos resultados satisfatórios. Para isso, o registro dos espectros das amostras de semente dos feijões foram medidos em ambos os lados. Ou seja, as medidas das 90 amostras de sementes de cada cultivar geraram 180 espectros, totalizando 540 espectros das três cultivares de feijões. Uma média dos seis espectros, total das medidas dos dois lados do feijão, foi calculada para representar a amostra e, então, ser utilizada no procedimento quimiométrico.

5.2.3. Análise Quimiométrica

A capacidade discriminatória da técnica do espectrômetro portátil DLP NIRscan Nano EVM aliada às técnicas de reconhecimento de padrões, utilizaram as informações contidas nos espectros para a classificação das sementes das diferentes cultivares (BRSFC401RMD, BRSPontal e BRSEstilo) de feijões.

Inicialmente, uma análise exploratória dos dados utilizando a PCA foi realizada. O intuito foi de observar o comportamento das amostras no sentido de agrupamentos, amostras anômalas e a verificação de quais tipos de pré-processamentos seriam mais adequados para discriminação das classes.

De modo a garantir uma melhor representatividade dos modelos de classificação, as amostras foram divididas em três subconjuntos: treinamento (60%), validação (20%) e teste (20%) utilizando o algoritmo KENNARD-STONES (KS). Este algoritmo foi aplicado separadamente para cada classe, onde primeiramente as amostras do conjunto de treinamento foram inicialmente selecionados e os remanescentes foram divididos em validação e teste, de acordo com a ordenação do KS.

A **Tabela 5.2** apresenta o número de amostras de treinamento, validação e teste para as classes de sementes das cultivares (BRSFC401RMD, BRSPontal e BRSEstilo) de feijões.

Tabela 5.2. Número de amostras de treinamento, validação e teste para as classes de sementes de feijões

Classes	Subconjuntos		
	Treinamento	Validação	Teste
BRSFC401RMD	60	15	15
Transgênicos			
BRSPontal	60	15	15
Convencional			
BRSEstilo	60	15	15

As amostras foram escolhidas de modo que o subconjunto de treinamento possuisse a faixa mais ampla de espaço amostral e o conjunto teste, a mais estreita. Este cuidado é necessário, pois normalmente os modelos multivariados não podem ser usados para estimar as propriedades de interesse fora do domínio que o modelo foi desenvolvido. As amostras de treinamento e validação foram utilizadas para o procedimento de modelagem (incluindo a seleção de variáveis para os modelos LDA e a determinação do número de variáveis latentes para os modelos PLS-DA). Já o conjunto externo de teste foi utilizado apenas para uma avaliação final dos modelos de classificação (ACO-LDA, SPA-LDA, GA-LDA e SPA-DA). O conjunto de teste não participou da etapa da construção do modelo e foi

utilizado com a finalidade de comparar o desempenho de classificação dos diferentes modelos em função da taxa de classificação correta (TCC).

Os resultados obtidos do ACO-LDA foram comparados a outros métodos de seleção de variáveis bem estabelecidos na literatura como a Análise Discriminante pelos Mínimos Quadrados Parciais (PLS-DA), Algoritmo genético (GA) e o Algoritmo das Projeções Sucessivas (SPA), ambos associados à modelagem LDA.

O pré-processamento dos dados e a PCA foram realizados utilizando o programa The Unscrambler® X.1 (Camo/SA). O algoritmo Kennard-Stonne (KS), ACO, GA, SPA e PLS-DA foram executados utilizando o programa Matlab®7.10.

5.3. Resultados e Discussão

5.3.1. Perfil espectral das amostras de feijão

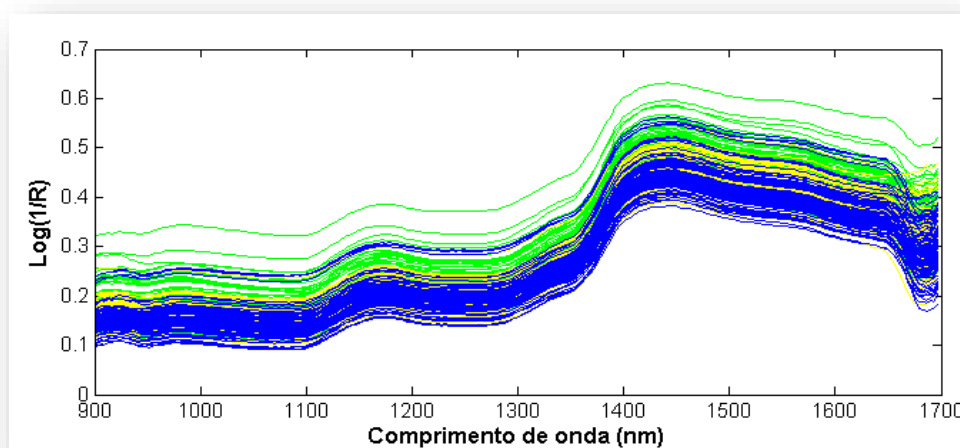


Figura 5.9. Espectros originais na região do infravermelho próximo das sementes dos feijões BRSFC401RMD (espectro verde), BRSPontal (espectro amarelo) e BRSEstilo (espectro azul).

A **Figura 5.9** mostra os espectros brutos de reflectância das sementes de feijões na região do infravermelho próximo registrados na faixa de 900 a 1700nm.

Como pode ser visto na **Figura 5.9**, as principais bandas nos espectros NIR das sementes de feijão estão nas faixas de 1100-1200 e

1400-1500 nm. O primeiro está associado ao segundo sobretudo da ligação C-H, enquanto o segundo está associado ao primeiro sobretudo de O-H e N-H, assim como as bandas de combinação de C-H. Além disso, outras bandas de menores magnitude são encontradas no espectro.

Observa-se que existe uma separação entre as duas classes (BRSFC401RMD e BRS Pontal) e uma sobreposição entre a classe BRSEstilo e BRSPontal. Todos esses fatores impossibilitam a distinção entre as classes através de uma inspeção visual dos espectros. Estes apresentam dispersão das classes associada às variações da linha de base, o que pode ser atribuído aos efeitos de espalhamento das amostras.

Na **Figura 5.10** pode-se observar o gráfico dos escores (PC1 x PC2) resultante da aplicação da PCA para os espectros brutos do NIR. Como pode ser visto, há uma aglomeração e sobreposição das cultivares.

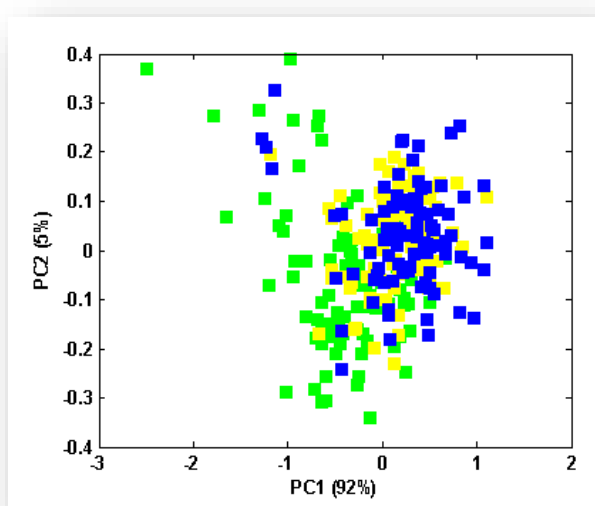


Figura 5.10. Gráfico dos escores PC1 x PC2 para as sementes de feijões na região do NIR. (quadrado **Verde**: BRSFC401RMD;quadrado **Amarelo**: BRSPontal; quadrado **Azul**: BRSEstilo)

Torna-se necessário, então, avaliar o uso de diferentes tipos de pré-processamentos dos dados. A **Figura 5.11** apresenta os espectros com sete diferentes tipos de pré-processamentos.

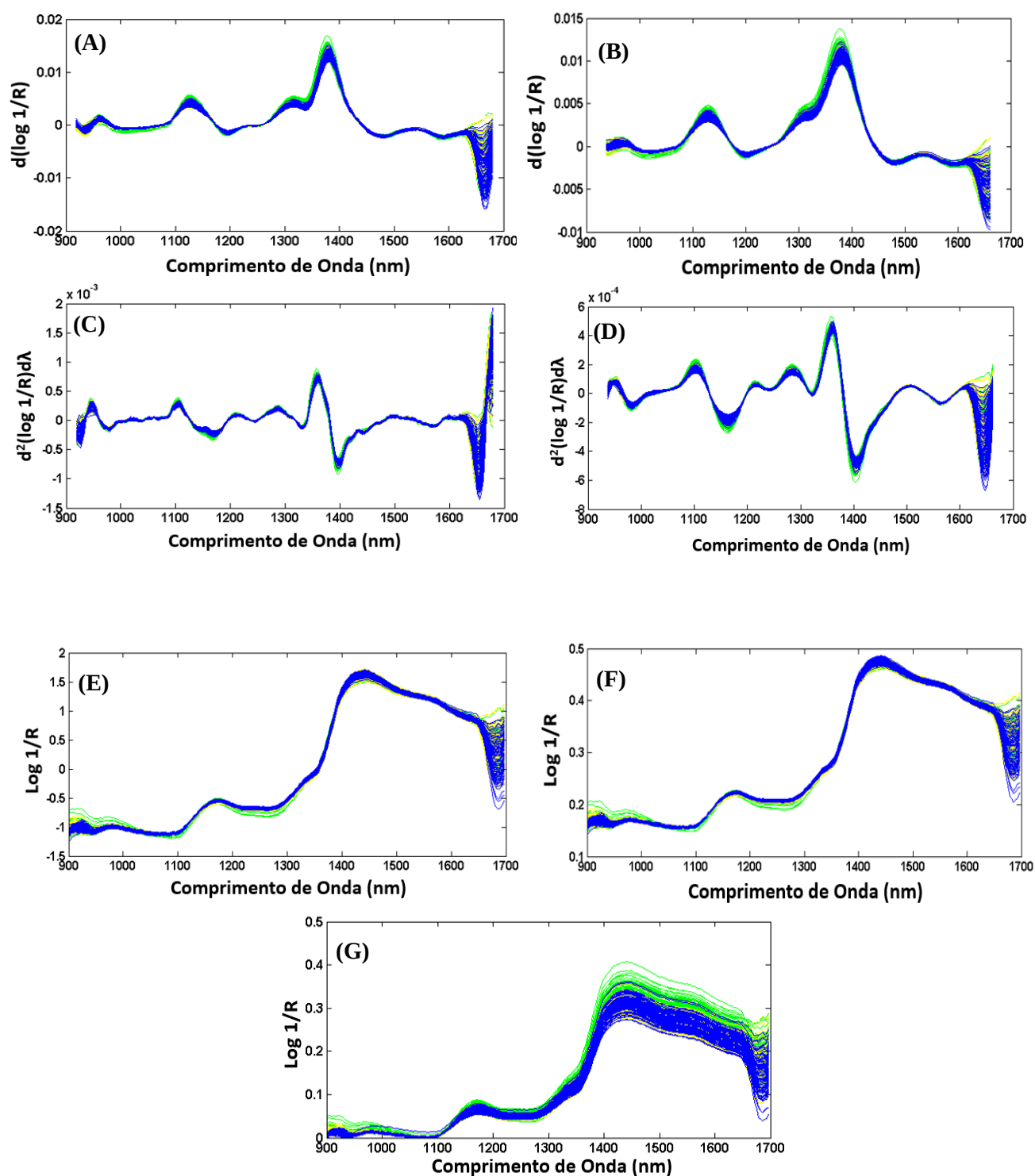
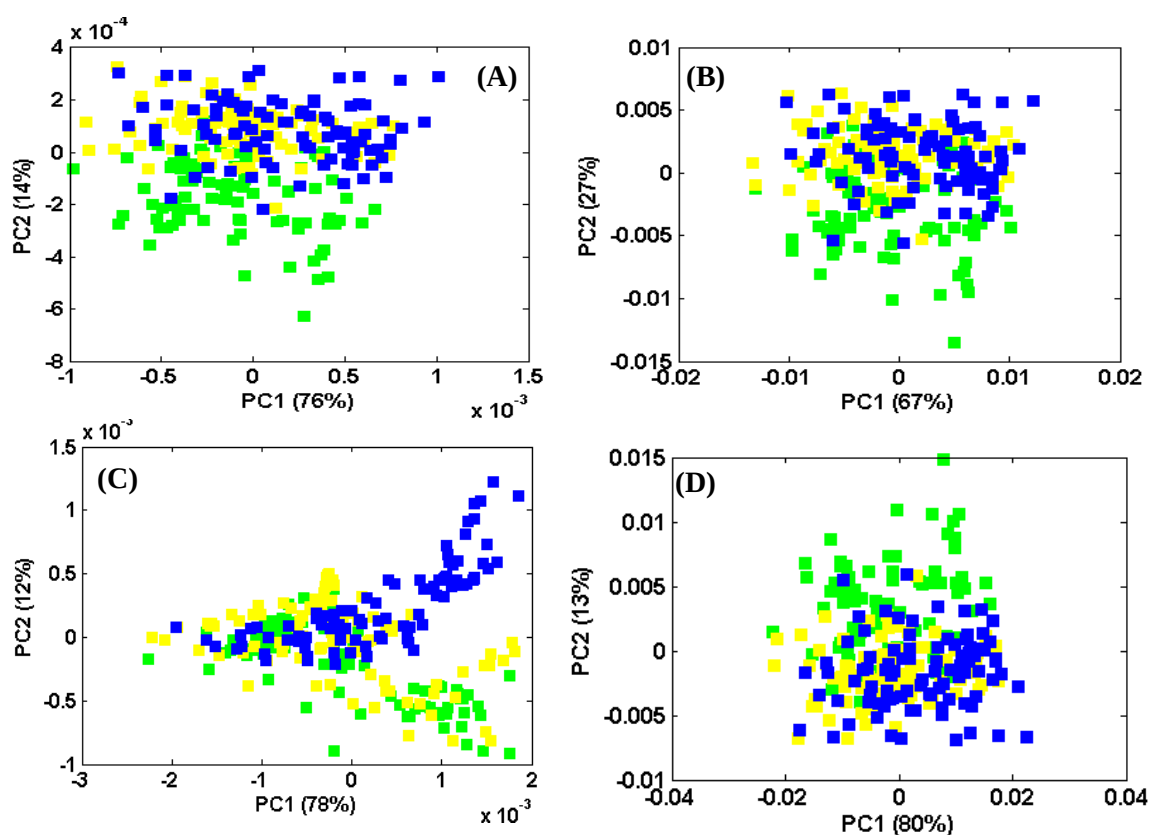


Figura 5.11. Espectros pré-processados com 1º derivada utilizando o algoritmo S.G com polinômio e janela de: A) 11 pontos e B) 21 pontos; 2º derivada utilizando o algoritmo S.G com polinômio e janela de C) 11 pontos e D) 21 pontos; E) MSC; F) SNV; G) Baseline. (Espectro Verde: BRSFC401RMD; espectro Amarelo: BRSPontal; espectro Azul: BRSEstilo)

Com base nos gráficos apresentados nas **Figuras 5.11** é possível observar que os pré-processamentos empregados corrigiram a variação sistemática da linha de base, bem como o espalhamento das amostras. Porém, ainda não é possível identificar as diferentes classes através de uma inspeção visual dos espectros. Assim, a aplicação de métodos de reconhecimento de padrão torna-se necessário.

5.3.2. Análise por Componentes Principais (PCA)

Com a finalidade de realizar um estudo exploratório das cultivares de semente de feijões, uma PCA foi aplicada aos espectros pré-processados. Um modelo, com a seleção de todas variáveis espectrais, utilizando duas componentes principais e as variâncias dos dados explicadas foram obtidos. As **Figuras 5.12** mostram o gráfico dos escores obtidos por PC2 *versus* PC1 de todas as amostras.



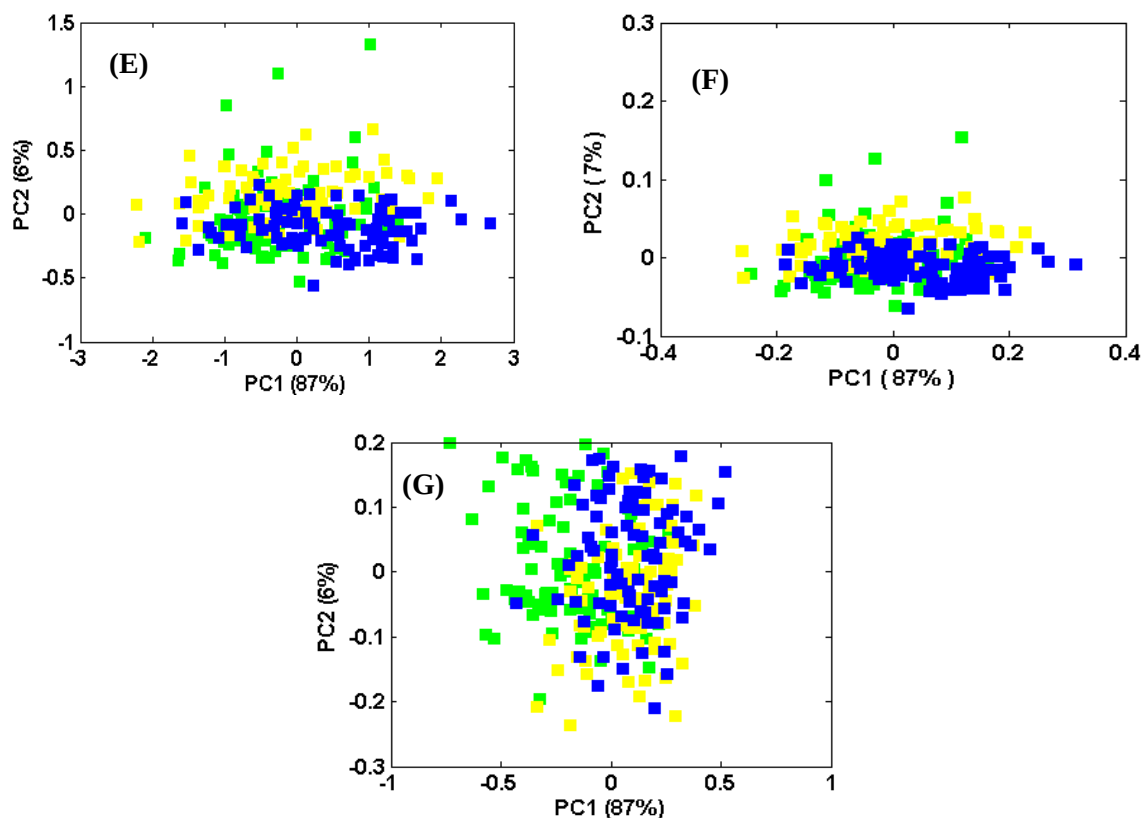


Figura 5.12. PCA para as sementes de feijões na região do NIR na faixa de 900 a 1700. A) 1º Savitzky-Golay (11); B) 1º Savitzky-Golay (21); C) 2º Savitzky-Golay (11); D) 2º Savitzky-Golay (21) E) MSC; F) SNV; G) Baseline. (**Verde:** BRSFC401RMD; **Amarelo:** BRSPontal; **Azul:** BRSEstilo)

Os gráficos dos escores obtidos por PC2 × PC1 mostram uma tendência de sobreposições entre as três classes das sementes de feijões estudadas nesse trabalho. Este comportamento condiz com o perfil espectral dessas amostras, uma vez que poucas diferenças são observadas. Diante do exposto, fez-se necessário a retirada da região que não informativa, e passando a analisar a faixa de 900 a 1635 nm. Então um novo estudo exploratório dos dados foram realizados conforme mostrado nas **Figuras 5.13**.

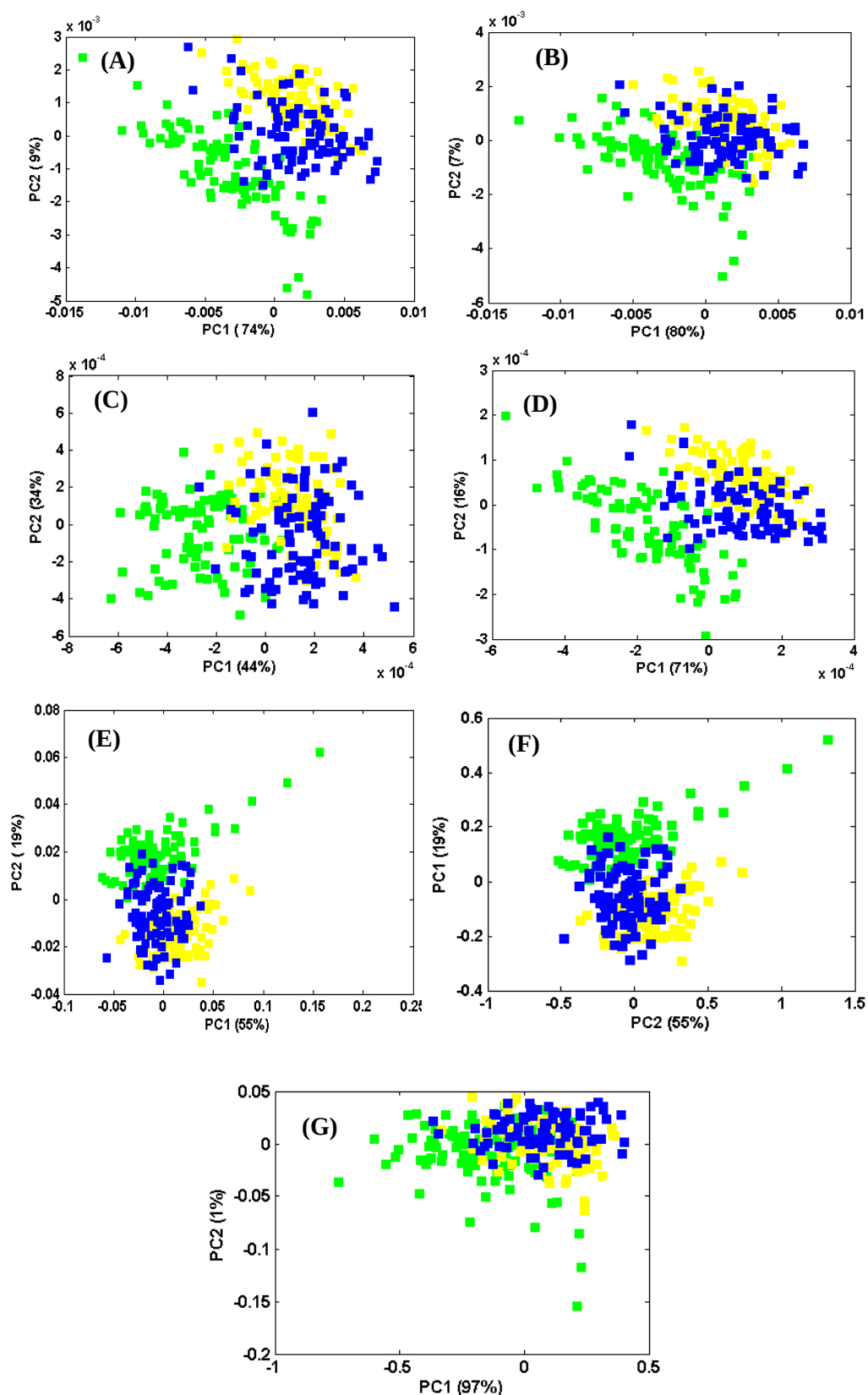


Figura 5.13. PCA para as sementes de feijões na região do NIR de 930 a 1630. A) 1^o Savitzky-Golay (11); B) 1^o Savitzky-Golay (21); C) 2^o Savitzky-Golay (11); D) 2^o Savitzky-Golay (21) E) MSC; F) SNV; G) Baseline. (**Verde:** BRSFC401RMD; **Amarelo:** BRSPontal; **Azul:** BRSEstilo)

Os gráficos dos escores (**Figura 5.13**) obtidos mostraram uma forte tendência de separação, nos espectros derivativos, entre dois grandes grupos conforme observado em **A**, **B**, **C** e **D**. Esta conclusão indica que as diferenças espectrais entre os dois tipos de sementes, das três estudadas, embora sutis, são sistemáticas e podem ser usadas para fins de discriminação. Além disso, pode argumentar-se que estas diferenças não se limitam à dispersão, uma vez que estes efeitos físicos foram amplamente eliminados pela derivada. Enquanto nos outros três pré-processamento aplicados (**E** e **F**), pode-se também observar comportamento semelhante. E na **Figura 5.13. G**, não foi observado uma separação dos grupos pela inspeção visual.

Para entendermos quais variáveis contribuíram para a maior discriminação entre os grupos, foi construído o gráfico dos pesos (**Figura 5.14**). Os pesos em PC1 têm um perfil similar aos espectros derivados na **Figura 5.11.b**, ou seja, todos os picos espectrais estão presentes nos pesos em PC1. Assim, pode-se concluir que todo o espectro contribui para a separação das classes. Este resultado é razoável, uma vez que os principais componentes químicos são os mesmos em ambos os tipos de sementes, com pequenas alterações na composição.

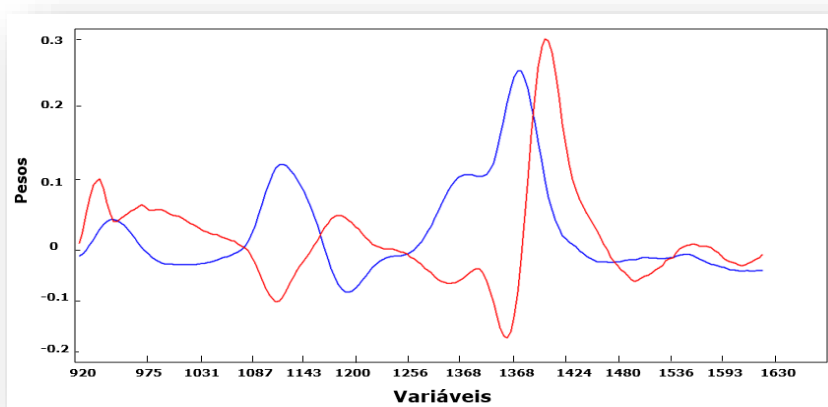


Figura 5.14. Gráficos dos pesos *versus* variáveis de PC1(azul) e PC2(vermelho) de todas as classes.

Tais resultados sugerem uma investigação com base no emprego de técnicas de seleção de variáveis e métodos de reconhecimento de padrões supervisionados com o intuito de melhorar a discriminação entre as diferentes amostras, bem como construir modelos capazes de realizar previsões em amostras futuras.

5.3.3. Métodos de Reconhecimento de Padrões Supervisionados

Os modelos LDA foram criados e validados utilizando cadeias de variáveis selecionadas baseadas nos parâmetros utilizados por cada algoritmo (SPA, ACO e GA). Tais modelos foram aplicados para a classificação do conjunto de teste. A **Tabela 5.3** apresenta a taxa de classificação correta obtida pelos modelos LDA-SPA, LDA-ACO, LDA-GA e para o conjunto de teste, com diferentes pré-processamentos.

Tabela 5.3. Taxa de classificação correta obtida pelos modelos de classificação LDA-ACO, LDA-GA e LDA-SPA aplicados aos dados do NIR para o conjunto de teste. Os valores em parênteses indicam o número de variáveis selecionados.

Taxa de Classificação Correta (%)			
Pré-Processamentos	LDA-ACO	LDA-GA	LDA-SPA
Sem pré-processamento	100 (10)	98.3 (13)	98.3 (10)
1º Derivada 11 pontos	100 (7)	98.3 (13)	98.3 (10)
1º Derivada 21 pontos	100 (10)	100 (11)	100 (10)
2º Derivada 11 pontos	100 (9)	98.3 (8)	100 (10)
2º Derivada 21 pontos	100 (9)	96.6 (9)	98.3 (6)
MSC	96.6 (7)	98.3 (8)	100 (9)
SNV	98 (10)	100 (9)	98.3 (8)
Baseline	98.3 (7)	96.6 (9)	100 (14)

De acordo com a **Tabela 5.3**, pode-se observar que a maioria dos resultados são bastantes semelhantes entre os pré-processamentos

aplicados, não observando diferença significativas entre eles. Diante disso, adotamos como os melhores resultados alcançados para o desenvolvimento do trabalho, foi utilizando a primeira derivada de 21 pontos com o pré-processamento que atingiu 100% de acertos no conjunto de teste para todos os métodos. De uma forma geral, os modelos desenvolvidos com uso de diferentes pré-processamentos apresentaram desempenhos muito parecidos.

5.3.3.1. ACO-LDA

A otimização dos parâmetros de ACO que proporcionaram melhores resultados foram: taxa de evaporação de feromônio de 65%, 100 ciclos, colônias de 50 formigas, 20 variáveis e 30% de formigas cegas. A **Figura 5.15** mostra os espectros médios de cada classe de sementes de feijões do melhor resultado das 5 execuções testadas, resultando em um total de 10 comprimentos de onda (972, 1010,5, 1192,5, 1315, 1371, 1409,5, 1427, 1437,5, 1455 e 1525 nm) selecionados pelo ACO-LDA.

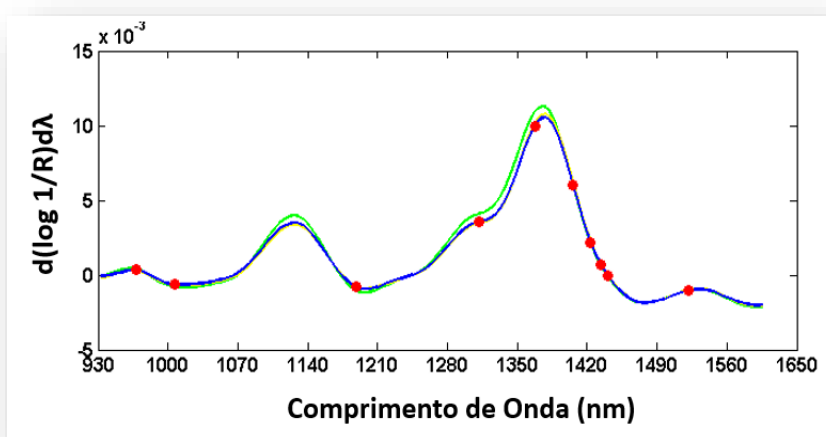


Figura 5.15. Espectro médio para cada tipo semente analisado. Os onze comprimentos de ondas selecionados pelo ACO-LDA encontram-se indicados em círculos.

As 10 variáveis selecionadas pelo ACO para o modelo LDA usando os dados NIR (**Figura 5.15**) encontram-se em várias partes das regiões dos espectros. Os comprimentos de ondas selecionados foram: 972 e 1010 nm

estão associados ao terceiro sobretom dos modos de estiramento das ligações C-H; 1193 nm ao segundo sobretom dos modos de estiramento das ligações C-H; 1315, 1371, 1409, 1427, 1438, 1455 e 1525 nm ao primeiro sobretom das combinação de ligações C-H. E estão associado ao primeiro sobretom de N-H das proteínas, a variáveis selecionadas entre a faixa de 1400 a 1500 nm.

O modelo ACO-LDA construído com essas variáveis alcançaram uma TCC de 99,7% e 100 % para o conjunto de treinamento e validação respectivamente.

Por meio de uma inspeção visual realizada nos gráfico dos escores (Figura 5.16), é possível verificar a capacidade discriminante do método avaliado para as sementes das três classes de feijões.

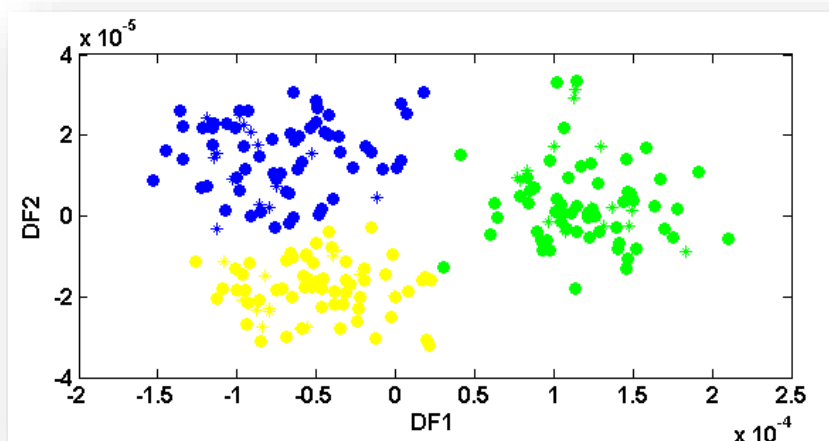


Figura: 5.16. Gráfico dos escores da função discriminante 1 para a espectroscopia NIR das sementes BRSFC401RMD(●), BRSPontal (●) e BRSEstilo (●) obtidos pela ACO-LDA.

Como pode ser observado na Figura 5.16, as amostras de uma mesma classe tendem a agrupar-se, ocupando a mesma região no gráfico dos escores. Enquanto os cultivares (BRSPontal e BRSEstilo) estão mais à esquerda (escores negativos). O comportamento mostra que existe uma discriminação entre a cultivar das sementes de feijões BRSFC401RMD das outras cultivares (BRSPontal e BRSEstilo). Os resultados obtidos pela LDA com as variáveis selecionadas pelo ACO foram bastante efetivos na

discriminação das classes quando aplicados para a classificação do conjunto de teste, alcançando uma TCC de 100%.

Para corroborar com o gráfico dos escores, podemos observar o comportamento das variáveis selecionadas pelo ACO que melhor discrimina as classes no gráfico dos pesos (**Figura 5.17**).

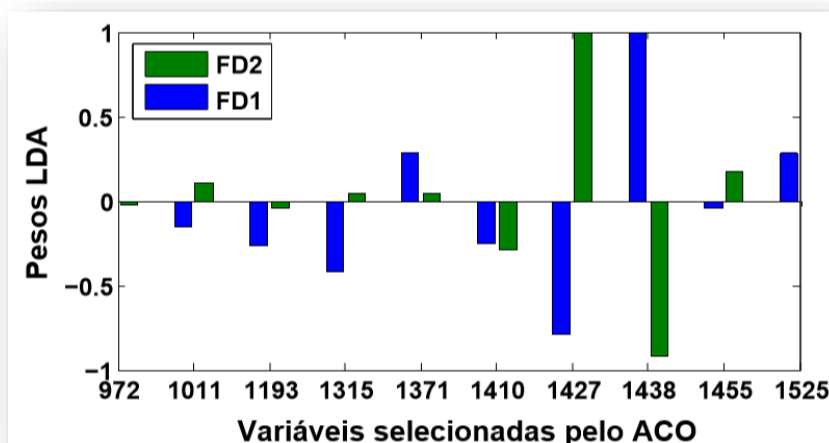


Figura 5.17. Gráfico dos pesos das variáveis selecionadas pelo ACO-LDA para as três cultivares dos feijões.

O gráfico dos pesos na **Figura 5.17**, mostra a intensidade de cada variáveis escolhidas para as funções discriminantes (FD1 e FD2). Para a FD1 e FD2, as variáveis de maiores pesos foram os comprimentos de onda de 1427 nm e 1438 nm. Cada um contribuindo positivamente e negativamente para a separação das classes de feijões. Esses comportamentos podem estar associados ao primeiro sobretom de ligações O-H e N-H, assim como as bandas de combinação de ligações C-H das três cultivares de feijão.

5.3.3.2. SPA-LDA

A **Figura 5.18** mostra os espectros médios de cada classe com os dez comprimentos de onda selecionados. Os comprimentos de onda selecionados foram: 937, 955 e 979 nm associado ao terceiro sobretom de estiramento da ligação C-H; 1063 ao estiramento e deformação da ligação C-H; 1130 nm ao segundo sobretom de estiramento da ligação C-H; 1371,

1406, 1441, 1522 e 1602 nm ao primeiro sobretom de combinação da ligação C-H.

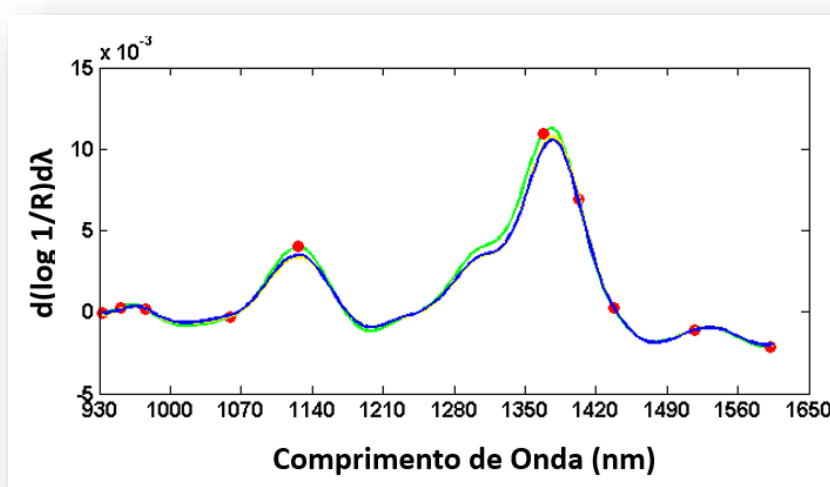


Figura 5.18. Espectro médio com as variáveis selecionadas pelo SPA

A LDA foi aplicada às variáveis selecionadas pelo SPA para o conjunto completo das amostras. É importante ressaltar que o modelo LDA gera um número de funções discriminantes igual ao número de classes menos um ($c - 1$). Portanto, neste trabalho, apenas uma função discriminante (FD1) foi gerada para as duas classes de semente. A **Figura 5.19** mostra o gráfico dos escores obtidos pela LDA aplicada ao conjunto completo de amostras usando as variáveis selecionadas pelo SPA.

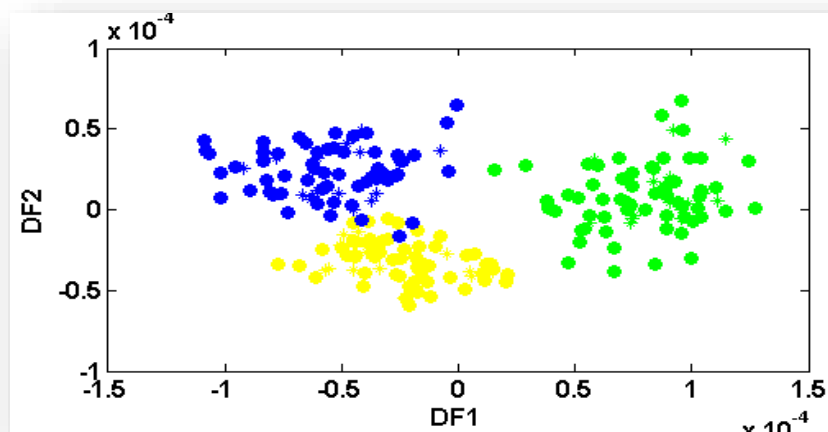


Figura 5.19. Gráfico dos escores da função discriminante 1 para a espectroscopia NIR das sementes BRSFC401RMD (●), BRS Pontal (●) e BRS Estilo (●) obtidos pela SPA-LD

Como pode ser observado na **Figura 5.18**, as amostras se comportam da forma semelhante ao ACO-LDA. Os resultados obtidos pela LDA com as variáveis selecionadas pelo SPA para a discriminação das classes quando aplicados a classificação do conjunto de teste, alcançaram uma TCC de 100%. Já a taxa de classificação correta para o conjunto de treinamento foi de 98,7%.

Para poder observar o comportamento das amostras distribuídas no gráfico dos escores, precisamos ter o conhecimento das variáveis que obtiveram os maiores pesos dentre as 10 variáveis selecionadas. A **Figura 5.20** mostra os pesos das variáveis selecionadas pelo SPA que melhor discrimina as classes.

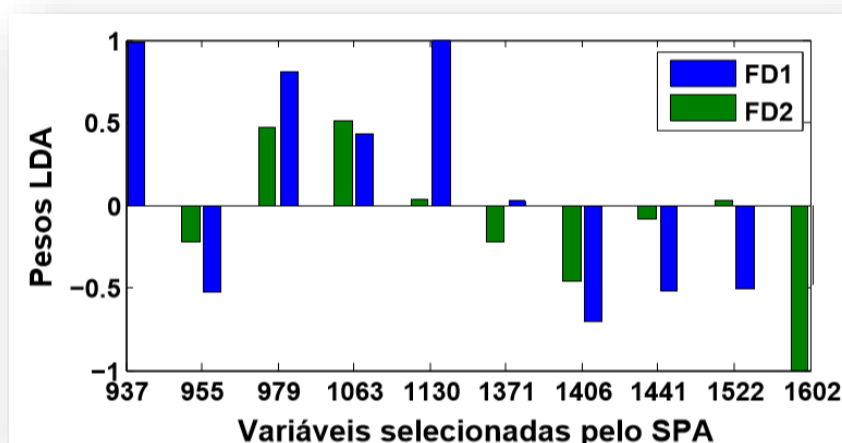


Figura 5.20. Gráfico dos pesos das variáveis selecionadas pelo SPA-LDA para as três cultivares dos feijões.

Como podemos observar na **Figura 5.20**, das duas funções discriminantes (FD1 e FD2), uma maior quantidade de variáveis com maiores pesos é observado em FD1. Ou seja, podem estar atribuídas a separação das classes convencionais dos transgênicos. Visto que esta se encontra na região positiva dos escores. O comportamento das variáveis de maiores pesos para FD1 foram os comprimentos de onda de 937, 979, 1130 e 1406nm. As duas primeiras estão associados ao terceiro sobretom de estiramento da ligação C-H e a terceira variável de maior peso ao segundo sobretom de estiramento da ligação C-H. Já a FD2 contribui com uma

variável (1602 nm) de maior peso, na região negativa, que podem estar relacionados ao primeiro sobretom de combinação da ligação C-H e assim a ocorrência da que separação das classes BRS pontal e BRS estilo.

5.3.3.3 GA-LDA

Para propósito de comparação, o GA também foi empregado para seleção de variáveis em LDA com os seguintes parâmetros: probabilidade de cruzamento de 60 e 100 gerações com população de 50 indivíduos. Além disso, o algoritmo foi repetido cinco vezes, a partir de populações diferentes. A melhor solução resultante dessas cinco realizações foi mantida e adotada como resultado do GA para este estudo de classificação.

A **Figura 5.21** mostra os espectros médios de cada classe de sementes de feijões analisada com a indicação dos onze comprimentos de onda selecionados pelo GA-LDA. As variáveis selecionadas foram: 937 e 958 nm associado ao terceiro sobretom de estiramento da ligação C-H; 1094 nm ao estiramento e deformação da ligação C-H; 1192, 1210 e 1231 nm ao segundo sobretom de estiramento da ligação C-H; 1300, 1400, 1423, 1437 e 1595 nm ao primeiro sobretom de combinação da ligação C-H.

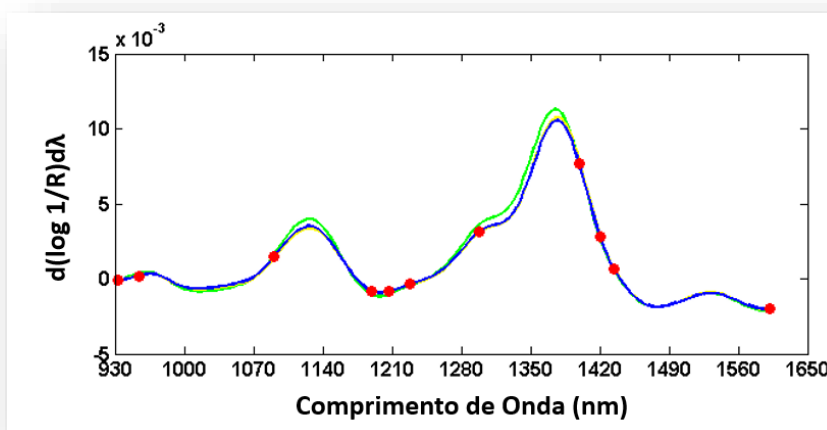


Figura 5.21. Espectro médio para cada tipo semente analisado. (a)transgênicos, (b) convencionais. Os seis comprimentos de ondas selecionados pelo GA-LDA encontram-se indicados em círculos.

O modelo LDA resultante construído com as onze variáveis selecionadas pelo GA, quando aplicado ao conjunto de treinamento, apresentou um desempenho similar ao SPA-LDA (três erros de classificação). A taxa de classificação correta para o conjunto de treinamento foi de 95.8%. Na **Figura 5.22** é exposto o gráfico dos escores obtidos pela LDA para o conjunto completo de amostras.

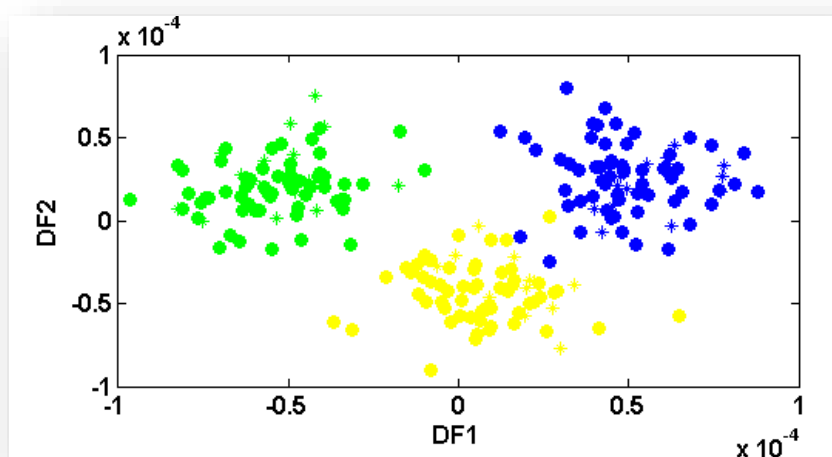


Figura 5.22. Gráfico dos escores da função discriminante 1 para a espectroscopia NIR das sementes BRSFC401RMD (●), BRSPontal (●) e BRSRstilo (●) obtidos pela GA-LDA.

Observa-se que as amostras BRSFC401RMD apresentaram valores de escores mais negativos de DF1, quando comparadas com as outras duas classes. Onde estas, estão alocadas em sua maioria nos escores positivos. Os resultados obtidos pela LDA com as variáveis selecionadas pelo GA para a discriminação das classes quando aplicados a classificação do conjunto de teste, alcançaram uma TCC de 100%.

Na **Figura 5.23**, é mostrado o comportamento dos pesos das variáveis selecionadas pelo GA. E pode-se perceber que na DF1 as variáveis de maior peso foram dois comprimentos de onda bem próximos: 1424 nm, contribuindo positivamente e 1438 nm, contribuindo negativamente. Isso pode estar atribuído ao primeiro sobretom de combinação da ligação C-H.

E os pesos da FD2 mais intenso foi: 1210 nm, contribuindo positivamente associado ao segundo sobretom de estiramento da ligação C-H.

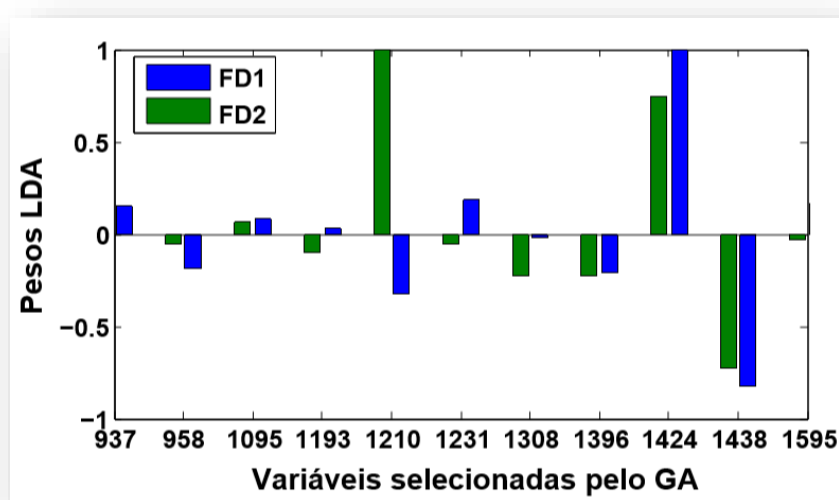


Figura 5.23. Gráfico dos pesos das variáveis selecionadas pelo SPA-LDA para as três cultivares dos feijões.

5.3.3.4. PLS-DA

O modelo PLS-DA foi construído para as três classes de feijões estudadas usando conjunto externo de validação (validation set). O número de fatores ótimos a ser usado no modelo foi determinado com base na função da taxa de erro. E esse número de fatores ideal para a construção do modelo PLS-DA, com a menor taxa de erros, foram de Quatorze fatores. O perfil da taxa de erro em virtude da inclusão de fatores ao modelo é apresentado na **Figura 5.24.**

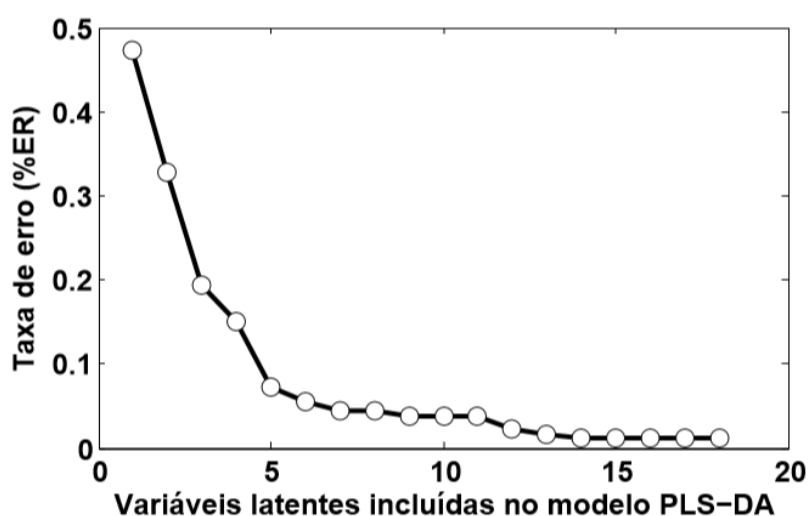


Figura 5.24: Taxa de erro versus o número de variáveis latentes incluídas no modelo.

A partir dos resultados apresentados na **Tabela 5.4**, são observados as informações obtidas pelos modelos ACO-LDA, GA-LDA, SPA-LDA e PLS-DA para o conjunto de treinamento das amostras de feijões. Como podemos ver, todos os modelos obtiveram um pequeno desempenho abaixo de 100% em termos de ajuste. O modelo ACO-LDA fez uma classificação errada na etapa de validação. Uma amostra da classe (Pontal) foi designada como membro da classe 3 (Estilo), obtendo um desempenho de 99.6%. No modelo GA-LDA, uma amostra da classe 2 (Pontal) foi classificada como sendo da classe 3 (Estilo) e duas amostras da classe 3 (Estilo) como sendo da classe 2 (Pontal), obtendo um desempenho de 98.7%. E no modelo SPA-LDA o erro de classificação já foi um pouco maior. Onde três amostras da classe 2 (Pontal) foram classificadas como sendo da classe 3 (Estilo) e cinco amostras da classe 3 (Estilo) como sendo da classe 2 (Pontal), obtendo um desempenho de 96.6%. E por fim, o PLS-DA teve a mesma quantidade de erros do ACO-LDA. Portanto, uma comparação entre ACO, GA e SPA mostram resultados semelhantes em termos de precisão, bem como o PLS-DA. Note que o ACO selecionou apenas 10 variáveis, O GA selecionou 11 variáveis, SPA selecionou também 10 variáveis e o PLS-DA 14 variáveis latentes. Dentre os modelos LDA, o

ACO foi o que obteve o menor erro de classificação. E comparando o ACO com o PLS-DA, foi o que obteve a menor quantidade de variáveis. Diante disso, o ACO é um algoritmo promissor na seleção de variáveis informativas das amostras de feijões para a classificação das classes a partir de dados espectrométricos NIR.

Tabela 5.4 Resultados de previsão para um conjunto de treinamento: matriz de confusão, sensibilidade e especificidade.

		Treinamento		
		Transgênico	Pontal	Estilo
ACO-LDA	Transgênico	60	-	-
	Pontal	-	59	1
	Estilo	-	-	60
Métrica ACO-LDA	Sensibilidade	1	0.98	1
	Especificidade	1	1	0.99
	ACC (%)		99.6	
GA-LDA	Transgênico	60	-	-
	Pontal	-	59	1
	Estilo	-	2	58
Métrica GA-LDA	Sensibilidade	1	0.98	0.96
	Especificidade	1	0.98	0.99
	ACC (%)		98.7	
PLS-LDA (14 latent variables) ^a	Transgênico	60	-	-
	Pontal	-	59	1
	Estilo	-	-	60
Métrica PLS-DA	Sensibilidade	1	0.98	-
	Especificidade	1	1	0.99
	ACC (%)		99	
SPA-LDA	Transgênico	60	-	-
	Pontal	-	57	3
	Estilo	-	5	55
Métrica SPA-LDA	Sensibilidade	-	0.95	0.916
	Especificidade	1	0.958	0.975
	ACC (%)		96.6	

Através dos resultados da **Tabela 5.4**, é possível determinar a sensibilidade e a especificidade do modelo construído. No modelo construído a sensibilidade para a classe de Estilo foi igual a 1 enquanto que para a classe pontal foi igual a 0.98. Isso significa que o modelo não foi capaz de classificar corretamente todas as amostras da classe pontal. Na classe estilo a especificidade foi igual a 0.99, pois uma amostra de validação da

classe pontal foram prevista na classe estilo. Por outro lado, a especificidade para a classe pontal foi igual a 1 devido a nenhuma amostra de validação da classe pontal ter sido prevista como sendo pertencente à classe estilo.

Os resultados na etapa de validação não foram apresentadas porque todos os modelos apresentaram 100% de acertos.

A **Tabela 5.5** apresenta o resumo dos erros de classificação para o conjunto das amostras de previsão das quatro estratégias de modelagem empregadas nesse estudo.

Tabela 5.5 Resultados de previsão para um conjunto de teste: matriz de confusão, sensibilidade e especificidade

Classe Verdadeira	ACO-LDA			GA-LDA			PLS-DA			SPA-LDA		
	Classe Predita			Classe Predita			Classe Predita			Classe Predita		
	T	P	E	T	P	E	T	P	E	T	P	E
T (15)	15	-	-	15	-	-	15	-	-	15	-	-
P (15)	-	15	-	-	15	-	-	15	-	-	15	-
T (15)	-	-	15	-	-	15	-	-	15	-	-	15
Sensibilidade	1	1	1	1	1	1	1	1	1	1	1	1
Especificidade	1	1	1	1	1	1	1	1	1	1	1	1
TCC (%)	100			100			100			100		

T(Transgênico), P (Pontal), E (Estilo)

Ao reforçar os resultados obtidos nas etapas de ajuste e validação, podemos ver que, com base nos resultados apresentados na **Tabela 5.5**, observa-se que todos os modelos obtiveram sucesso na classificação de um conjunto de amostras externas. A precisão obtida foi 100% para todos os modelos ACO-LDA, GA-LDA, SPA-LDA e PLS-DA, respectivamente. Esses resultados indicam que o ACO, quando combinado com o LDA, é capaz de fornecer resultados compatíveis com aqueles obtidos por ferramentas quimiométricas poderosas e versáteis, como SPA e GA.

CAPÍTULO VI

CONCLUSÕES E PERSPECTIVAS FUTURAS



6.1 CONCLUSÕES

Um novo algoritmo para resolver problemas de classificação, baseado na otimização de colônias de formigas para seleção de variáveis acopladas à análise discriminante linear, foi demonstrado neste trabalho. O desenvolvimento dessa abordagem de seleção de variáveis envolve uma meta-heurística que imita o comportamento cooperativo da formiga para encontrar um pequeno subconjunto capaz de produzir modelos ACO-LDA satisfatórios.

A cooperação do tipo formiga leva ao feromônio depositado aleatoriamente no elemento vetorial correspondente às variáveis mais relevantes, aumentando a probabilidade de descartar variáveis não informativas e de escolher as mais relevantes. Essa característica, ausente no algoritmo genético, favorece a construção de modelos ACO-LDA mais parcimoniosos do que é possível com o uso do algoritmo genético - análise de discriminação linear (GA-LDA).

O algoritmo ACO-LDA proposto foi avaliado em três problemas de classificação envolvendo duas diferentes técnicas analíticas (espectrometria NIR e UV-VIS) nos seguintes estudos de caso: (i) classificação de óleos vegetais comestíveis via espectrometria UV-VIS, (ii) classificação de amostras de chá em relação ao tipo e origem geográfica via espectroscopia NIR e (iii) Classificação entre três cultivares de feijões baseada em medidas nanoNIR. Nos três casos, o método ACO-LDA selecionou um pequeno subconjunto de variáveis, o que levou à classificação correta de quase todas as amostras e obtendo resultados consideravelmente melhores GA-LDA e tão bons quanto aqueles encontrados usando PLS-DA, ferramentas quimiométricas poderosas e bem estabelecidas.

Nas duas primeiras aplicações, a taxa de classificação correta de previsão foram: 100%, 95,7% e 100% dos modelos ACO-LDA, GA-LDA e PLS-DA respectivamente para classificação das amostras de óleos vegetais (UV-Vis). Enquanto para classificação das amostras de chás (NIR), a taxa

de classificação correta foram de 92%, 84% e 100% dos modelos ACO-LDA, GA-LDA e PLS-DA respectivamente. Já na terceira aplicação, a taxa de classificação correta foi de 100% para os quatros modelos (ACO-LDA, GA-LDA, SPA-LDA e PLS-DA) na classificação das amostras de feijões. Assim, o ACO acoplado ao LDA minimizou os problemas de generalização encontrados nos modelos de classificação de óleos vegetais, chás e feijões, os quais estão associados à multicolinearidade frequentemente presente nos dados de UV e NIR.

Portanto, O ACO-LDA pode ser considerado um método muito promissor e eficaz para seleção de variáveis em modelagem multivariada aplicada a problemas de classificação.

6.2 PERSPECTIVAS FUTURAS

- ✚ Construir gráficos de frequências das variáveis selecionadas nos ciclos de Monte Carlo para o conjunto de dados dos feijões, visando avaliar a convergência do algoritmo para as mesmas soluções.
- ✚ No estudo de classificação dos feijões, pretende-se incluir mais uma classe de transgênico.
- ✚ Avaliar o potencial do ACO-LDA em outros problemas de classificação e em outras técnicas analíticas.
- ✚ Adaptar o ACO para realizar classificação usando metodologias de classificação não linear.

REFERÊNCIAS

1. M. I. S. Gonçalves, W. T. S. Vilar, E. P. Medeiros, M. J. C. Pontes. A novel strategy for the classification of naturally colored cotton fibers based on digital imaging and pattern recognition techniques. *Analytical Methods*, 7 (2015) 5869–5875.
2. V. Boeris, J. A. Arancibia, A. C. Olivieri, Determination of five pesticides in juice, fruit and vegetable samples by means of liquid chromatography combined with multivariate curve resolution, *Analytica Chimica Acta* 814 (2014) 23-30.
3. L. R. Brito, M. P.F. Silva, J. J. R. Rohwedder, C. Pasquini, F. A. Honorato, M. F. Pimentel, Determination of detergent and dispensant additives in gasoline by ring-oven and near infrared hyperspectral imaging, *Analytica Chimica Acta* 863 (2015) 9-19.
4. G. D. Pierini, M. F. Pistonesi, M. S. Di Nezio, M. E. Centurión. A pencil-lead bismuth film electrode and chemometric tools for simultaneous determination of heavy metals in propolis samples, *Microchemical Journal*, 125 (2016) 266-272.
5. I. Lavilla, V. Romero, I. Costa, C. Bendicho. Greener derivatization in analytical chemistry. *Trends Analytica chemistry*. 61 (2014) 1-10.
6. M. Kamruzzaman, Y. Makino, S. Oshita, Non-invasive analytical technology for the detection of contamination, adulteration, and authenticity of meat, poultry, and fish: A review, *Analytica Chimica Acta* 853 (2015) 19-29.
7. S. Garrigues, M. Guardia, Non-invasive analysis of solid samples, *TrAC Trends in Analytical Chemistry* 43 (2013) 161-173.
8. U. T. C. P. Souto, M. F. Barbosa, H. V. Dantas, A. S. Pontes, W. S. Lyra, P. H. G. D. Diniz, M. C. U. Araújo, E. C. Silva. Identification of adulteration in ground roasted coffees using UV–Vis spectroscopy and SPA-LDA. *Lebensmittel Wissenschaft Technologie - Food Science e Technology*. 63 (2015) 1037-1041.

9. M. B. H. Santos, A. A. Gomes, W. T. S. Vilar, P. B. A. Almeida, M. Milani, M. B. M. Nóbrega, E. P. Medeiros, R. K. H. Galvão, M. C. U. Araújo. Non-Destructive NIR Spectrometric Cultivar Discrimination of Castor Seeds Resulting from Breeding Programs. *Journal of the Brazilian Chemical Society*. 25 (2014) 969-974.
10. S. M. Azcarate, A. A. Gomes, M. R. Alcaraz, M. C. U. Araújo, J. M. Camiña, H. C. Goicoechea, Modeling excitation - emission fluorescence matrices with pattern recognition algorithms for classification of Argentine white wines according grape variety, *Food Chemistry* 184 (2015) 214-219.
11. F. F. Gambarra-Neto, G. Marino, M. C.U. Araújo, R.K.H. Galvão, M.J.C. Pontes, E. P. de Medeiros, R.S. Lima. Classification of edible vegetable oils using square wave voltammetry with multivariate data analysis. *Talanta*. 77 (2009) 1660-1666.
12. D. Ballabio, V. Consonni, A. Mauri, R. Todeschini, Canonical Measure of Correlation (CMC) and Canonical Measure of Distance (CMD) between sets of data. Part 3. Variable selection in classification, *Analytica Chimica Acta* 657 (2010) 116-122.
13. P. Oliveri, Class-modelling in food analytical chemistry: Development, sampling, optimisation and validation issues - A tutorial, *Anal. Chim. Acta* 982 (2017) 9-19.
14. M. Barker, W. Rayens, Partial least squares for discrimination, *J. Chemom.* 17 (2003) 166-173.
15. C. D. Brown, R. L. Green, Critical factors limiting the interpretation of regression vectors in multivariate calibration, *Trends Analyt Chem.* 28 (2009) 506-514.
16. P. G. A. Nunes. Uma nova técnica para seleção de variáveis em calibração multivariada aplicada às espectrometrias uv-vis e nir. 2008. Tese (Doutorado em Química)-Curso de Pós Graduação em Química, Universidade Federal da Paraíba.

17. M. B. Seasholtz, B. Kowalski, The parsimony principle applied to multivariate calibration, *Analytica Chimica Acta* 277 (1993) 165-177.
18. J. Ghasemi, A. Niazi, R. Leardi. Geneti-algorihm-based wavelength selection in multicomponent spectrophotometric determination by PLS: application on copper and zinc mixture. *Talanta* 59 (2013) 311-317.
19. M. Forina, S. Lanteri, M. Casale, M. C. C. Oliveros. Stepwise orthogonalization of predictors in classification and regression techniques: An "old" technique revisited, *Chemometrics and Intelligent Laboratory Systems*, 87:252, 2007.
20. M. J. C. Pontes, R.K.H. Galvão, M. C. U. Araújo, P. N. T. Moreira, O. D. P. Neto, G. E. José, T. C. B. Saldanha, The successive projections algorithm for spectral variable selection in classification problems. *Chemometrics and Intelligent Laboratory Systems* 78: (2005) 11-18.
21. M. C. U Araújo, T.C.B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*. 57: (2001) 65-73.
22. Xue, B., Zhang, M. & Browne, W. N. Particle swarm optimisation for feature selection in classification : Novel initialisation and updating mechanisms. *Appl. Soft Comput.* 18, 261–276 (2014).
23. X. Wang, J. Yang, R. Jensen, X. Liu. Rough set feature selection and rule induction for prediction of malignancy degree in brain glioma. *Cumputer Methods and Programs in Biomedicine*. 83 (2006) 147-156.
24. S. Lin, S. Chen. PSOLDA : A Particle Swarm Optimization Approach for Enhancing Classification Accuracy Rate of Linear Discriminant Analysis. *Appl. Soft Comput.* 9 (2009) 1008–1015.
25. S. Sheykhizadeh, A. Naseri. An efficient swarm intelligence approach to feature selection based on invasive weed optimization: Application to multivariate calibration and classification using spectroscopic data. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* 194 (2018) 202-210.

26. M. Dorigo; V. Maniezzo; A. Colorni. Ant System: Optimization by a Colony of Cooperating Agents. IEEE Transaction of Systems, Man, and Cybernetics-Part B: Cybernetics. 26: (1996) 29-41.
27. O. Korb, T. Stützle, TE Exner, Uma abordagem de otimização de colônia de formigas para acoplamento flexível de ligante de proteína. Enxame Intelectual. 1 (2), 115–134 (2007)
28. T. L. A. Carvalho. Aplicação de algoritmos colônia de formigas na reconfiguração e alocação de bancos de capacitores em redes elétricas de distribuição. 2015. Dissertação (Mestrado em Engenharia Elétrica) Curso de Pós-Graduação em Engenharia Elétrica, Universidade Federal da Bahia.
29. M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond. Ant colony optimisation: a powerful tool for wavelength selection. J. Chemometrics 20 (2007) 146–157.
30. F. Allegrini, A. C. Olivieri. A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least - squares analysis. Analytica Chimica Acta 699: (2011) 18-25.
31. M. M. C. Ferreira, A. M. Antunes, M. S. Melgo, P. L. O. Volpe. Quimiometria I: calibração multivariada, um tutorial. Química Nova 22(1999) 274-731.
32. P. R. M. Correia, M. M. C. Ferreira, Reconhecimento de padrões por métodos não supervisionados: explorando procedimentos quimiométricos para tratamento de dados analíticos. Química Nova 30 (2007) 481–487.
33. K. S. Booksh, B. R. Kowalski, Theory of Analytical Chemistry, Analytical Chemistry, 66: (1994) 15, 1.
34. A. C. Olivieri, Recent advances in analytical calibration with multi-way data. Analytical. Methods 4 (2012) 1876- 1886.
35. G. M. Escandar, A. C. Oliveiri. Pratical three-way calibration. Elsevier. 2014.

36. [36] P. Valderrama, J. W.B. Braga, R. J. Poppi .Estado da arte de figuras de mérito em calibração multivariada: Revisão. *Química Nova* 32(2009) 1278-1287.
37. M. M. C. Ferreira. Multivariate QSAR. *Journal of Brazilian Chemical Society*, 13 (2002) 742-753.
38. L. A. Berrueta, R. M. Alonso-Salces, K. Héberger, Supervised pattern recognition in food analysis. *Journal of Chromatography A* 1158: (2007) 196–214.
39. D. A. Skoog, F. J. Holler, T. A. Nieman. *Princípios da análise instrumental*. 5.ed. Bookman, 2008.
40. A. A. Gomes. Algoritmo das Projeções Sucessivas aplicado à seleção de variáveis em regressão PLS. 2012. Dissertação (Mestrado em Química) Curso de Pós-Graduação em Química, Universidade do Federal da Paraíba.
41. M. J. C. Pontes. Classificação de bebidas alcoólicas destiladas e verificação de adulteração usando espectrometria NIR e quimiometria. João Pessoa, Programa de Pós-Graduação em Química, UFPB, 2005. Dissertação de Mestrado.
42. D. L. Massart, et al. *Journal Handbook of Chemometrics and Qualimetrics: Parte B*, Amsterdam: Elsevier, 1997.
43. A. F. Bueno. Desenvolvimento de um analisador de processo por espectroscopia no infravermelho próximo (NIR) para precisão de propriedades de derivados de petróleo. Campinas – SP, Programa de Pós Graduação em Química, UNICAMP, 2011, Tese de Doutorado, 264p.
44. R. A. G. Mendes. Utilização da espectroscopia em reflectância no infravermelho próximo para discriminação de espécies da família Myrtaceae. Brasília – DF, Programa de Pós Graduação em Biologia, UNB, 2014, Dissertação de Mestrado, 89 p.
45. R. J. Barnes, M. S. Dhanoa, S. J. Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, *Applied Spectroscopy*, 43:772,1989.

46. A. M. Souza, R. J. Poppi. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes Principais: um tutorial, parte I, *Química Nova*, 35 (2012) 223.
47. M. M. C. Ferreira. *Quimiometria: conceitos, métodos e aplicações*. 1.ed. Unicamp, 2016.
48. A. M. Souza, M. R. Coelho, P. Figueiras, T. A. F. Cunha, R. O. Dart, J. G. Parés, P. L. Simon, B. G. Cruz, R. J. Poppi, M. L. M. Santos, R. L. L. Berbara. Proposta de tutorial de Quimiometria utilizando técnicas modernas para a análise de solos. *Simpósio Brasileiro de Educação em Solos*, VI, 2012, Sobral-CE.
49. A. Savitzky, M. J. E. Golay. Smoothing and differentiation of data by simplified least-squares procedures, *Analytical Chemistry*, 8 (1964)1627.
50. A. G. González. Use and misuse of supervised pattern recognition methods for interpreting compositional data, *Journal of Chromatography A*, (2007) 1158:215.
51. M. J. C. Pontes. Algoritmo das Projeções Sucessivas para Seleção de Variáveis Espectrais em Problemas de Classificação. 2009.144p. Programa de Pós-Graduação em Química. (Tese de Doutorado) - Universidade Federal da Paraíba, João Pessoa.
52. A. Candolfi, R. D. Maesschalck, D. L. Massart, P. A. Hailey, A. C. E. Harrington, Identification of pharmaceutical excipients using NIR spectroscopy and SIMCA. *Journal of Pharmaceutical and Biomedical Analysis* 19: (1999) 923–935.
53. J. M. M. Neto, G. C. Moita. Uma introdução à análise exploratória de dados multivariados. *Química Nova* 21 (2008) 467-469.
54. P. R. M. Correia, M. M. C. Ferreira. Reconhecimento de padrões por métodos não supervisionados: explorando procedimentos quimiométricos para tratamento de dados analíticos. *Química Nova* 30: (2007) 481–487.

55. K.R. Beebe; R.J. Pell; M.B. Seasholtz. Chemometrics: a practical guide, New York: John Wiley and Sons, Inc., 1998.
56. K. D. T. M. Milanez. Classificação de óleos vegetais comestíveis usando imagens digitais e técnicas de reconhecimento de padrões. 2013. Dissertação (Mestrado em Química) - Curso de Pós-Graduação em Química, Universidade Federal da Paraíba.
57. L. Hansen. Análise de Dados Físico-Químicos de Amostras de Leite Cru do Sul do Brasil Utilizando Métodos Multivariados Exploratórios e Classificatórios. 2019. Dissertação (Mestrado em Química) - Curso de Pós-Graduação em Química, Universidade Federal do Rio Grande do Sul.
58. PEERBHAY, K.Y.; MUTANGA, O.; ISMAIL, R. Commercial tree species discrimination using airborne AISA Eagle hyperspectral imagery and partial least squares discriminant analysis (PLS-DA) in KwaZulu-Natal, South Africa, ISPRS Journal of Photogrammetry and Remote Sensing, 79: 19, 2013.
59. M.J.C. Pontes; C.F. Pereira; M.F. Pimentel; F.V.C. Vasconcelos; A.G.B. Silva. Screening analysis to detect adulteration in diesel/biodiesel blends using near infrared spectrometry and multivariate classification, Talanta, 85 (2011) 2159.
60. ALMEIDA, M.R.; CORREA, D.N.; ROCHA, W.F.C.; SCAFI, F.J.O.; POPPI, R.J. Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation, Microchemical Journal, 109 (2013) 170.
61. P.H.G.D Diniz. Novas estratégias para classificação simultânea do tipo e origem geográfica de chás. 2013. Tese (Doutorado em Química) - Curso de Pós-Graduação em Química, Universidade Federal da Paraíba.
62. L.A BARRUETA; R.M. ALONSO-SALCES;K. HÉBERGER. Supervised pattern recognition in food analysis. Journal of Chromatography A. 1158 (2007) 196-214.

63. BALLABIO, D., CONSONNI, V. Classification tools in chemistry. Part 1: linear models PLS-DA. *AnaliticalMethods* 5:3, 2013.
64. M. R. Almeida, D. N. Correa, W. F. C. Rocha, F. J. O. Scafi, R. J. Poppi. Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation, *Microchemical Journal*, 109 (2013) 170-177.
65. W. dos S. Gomes, F. L. Partelli, E. C. da S. Oliveira, et.al.. Espectroscopia de infravermelho para caracterização de genótipos de *coffea canephora*. *Ifesciencia* 6: 3 (2020) 35-44.
66. SILVA, A.C.; PONTES, L.F.B.L.; PIMENTEL, M.F.; PONTES, M.J.C. Detection of adulteration in hydrated ethyl alcohol fuel using infrared spectroscopy and supervised pattern recognition methods. *Talanta* 93: 129, 2012
67. R. A. Fisher. The use of multiple measurements in taxinomic problems, *Annales Eugenics*, 7:179, 1936.
68. K. Varmuza, P. Filzmoser.. *Introduction to multivariate statistical analysis in chemometrics*. (Taylor & Francis, 2008).
69. F. F. Gambarra-Neto. Classificação de óleos vegetais utilizando voltametria de onda quadrada e métodos quimiométricos. 2008. Dissertação (Mestrado em Química) Curso de Pós-Graduação em Química, Universidade Federal da Paraíba.
70. L. J. M. Coimbra, *et al.* Conseqüências da multicolinearidade sobre a análise de trilha em canola. *Ciência Rural* 35 (2005) 347–352.
71. A. V. Schenone, M. J. Culzoni, M. M. Galera, H.C. Goicoechea, Second-order advantage achieved by modeling excitation–emission fluorescence matrices affected by inner filter effects using a strategy which combines standardization and calibration: Reducing experimental and increasing analytical sensitivity. *Talanta* 109: (2013) 107–
72. A. C. Olivieri, N. M. Faber, J. Ferre, R. Boqué, J. H. Kalivas, M. Mark, Uncertainty estimation and figures of merit for multivariate calibration

- (IUPAC Technical Report). *Pure and Applied Chemistry* 78 (2006) 633-661.
73. A. C. Silva, S. F. C. Soares, M. Insausti, R. K. H. Galvão, B. S. Band, M. C. U. Araújo. Two-dimensional linear discriminant analysis for classification of three-way chemical data. *Analytica Chimica Acta* 938 (2016) 53-62.
 74. R. E. L. Abreu, J. E. M. Paz, A. C. Silva, M. J. C. Pontes, S. G. Lemos. Ethanol fuel adulteration with methanol assessed by cyclic voltammetry and multivariate calibration. *Fuel (Guildford)* 156 (2015) 20-25.
 75. M. N. Martins, R. K. H. Galvão, M. F. Pimentel. Multivariate Calibration Transfer Employing Variable Selection and Subagging. *Journal of the Brazilian Chemical Society (J. Braz. Chem. Soc.)* 21 (2010) 127-134.
 76. R. K. H. Galvão, S. F. C. Soares, M. N. Martins, M. F. Pimentel, Araújo, M. C. U. Araújo. Calibration transfer employing univariate correction and robust regression. *Analytica Chimica Acta* 864 (2015) 1-8.
 77. A.D. Walmsley. Improved variable selection procedure for multivariate linear regression. *Analytica Chimica Acta* 354 (1997) 225-232.
 78. R. N. F. Santos, R. K. H. Galvão, M. C. U. Araújo, E. C. Silva. Improvement of Prediction Ability of PLS Models Employing the Wavelet Packet Transform: A Case Study Concerning FT-IR Determination of Gasoline Parameters. *Talanta* 71 (2007) 1136-1143.
 79. A.A. Gomes, R.K.H. Galvão, M.C.U. Araújo, G. Vêras, E. C. Silva, The successive projections algorithm for interval selection in PLS. *Microchemical Journal* 100: (2013) 202-208.
 80. R. Leardi. Application of genetic algorithm-PLS for feature selection in spectral data sets. *Journal of Chemometrics*. 14 (2000) 643-655.
 81. X.Yang. Firefly algorithm, stochastic test functions and design optimisation. *Int. J. Bio-Inspired Comput.* 2 (2010) 78-84.
 82. M. C. U. Araújo, T. C. B. Saldanha, R. K. H. Galvão, T. Yoneyama, H. C. Chame, V. Visani. The successive projections algorithm for variable

- selection in spectroscopy multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems* 57 (2001) 65-73.
83. P. H. G. D. Diniz, M. F. Pistonesi, M. C. U. Araújo, Using iSPA-PLS and NIR spectroscopy for the determination of total polyphenols and moisture in commercial tea samples. *Analytical. Methods* 7 (2015) 3379-3384.
 84. M.J.C. Pontes, R.K.H. Galvão, M.C.U. Araújo, P.N.T. Moreira, O.D.P. Neto, G.E. José, T.C.B. Saldanha, The successive projections algorithm for spectral variable selection in classification problems. *Chemometrics and Intelligent Laboratory Systems* 78: (2005) 11-18.
 85. A. F. C. Pereira. Determinação simultânea de acidez, índice de refração e viscosidade em óleos vegetais usando espectrometria nir, calibração multivariada e seleção de variáveis. 2007. Programa de Pós-Graduação em Química - (Dissertação de Mestrado), Universidade Federal da Paraíba, João Pessoa.
 86. M. G. Varnamkhasti.; S. S Mohtasebi, M. L. R. Mendeza, A. A. Gomes, M. C. U. Araújo, R. K. H. Galvão, Screening analysis of beer ageing using near infrared spectroscopy and the Successive Projections Algorithm for variable selection. *Talanta* 89: (2012) 286-291.
 87. U. T. C. P. Souto, M. J. C. Pontes, E.C. Silva, R. K. H. Galvão, M. C. U. Araújo, F. A. C. Sanches, F. A. S. Cunha, M. S. R. Oliveira, UV-Vis spectrometric classification of coffees by SPA-LDA. *Food Chemistry* 119 (2010) 368-371.
 88. S. F. C. Soares, R. K. H. Galvão, M. J. C. Pontes, M. C. U. Araújo. A new validation criterion for guiding the selection of variables by the successive projections algorithm in classification problems, *Journal of the Brazilian Chemocal Society*. 25 (2014) 176-181.
 89. A.P. Weijer, et al. Using genetic algorithms for an artificial neural network model inversion. *Chemom. Int. lab. Sist.* 20:45,1993.
 90. P. H. de A, Konzen. et al. Otimização de métodos de controle de qualidade de fármacos usando algoritmo genético e busca de tabu. *Pesqui. Operacional* 23, (2003) 189-207 .

91. C. B. Lucasius, G. Kateman, Understanding and using genetic algorithms Part 1. Concepts, properties and context. *Chemometrics and Intelligent Laboratory Systems*. 19 (1993) 1-33.
92. C. A. Costa Filho, R. J. Poppi. Algoritmo Genético em química. *Química Nova* 22 (1999) 405.
93. R. Linden. *Algoritmos genéticos*. (Brasport livros e multimídia Ltda, 2008).
94. E. P., Bento, N. Kagan. Algoritmos genéticos e variantes na solução de configuração de redes de distribuição. *Rev. Control. Automação* 19, (2008) 302–315.
95. [95] B. Holldobler and E. O. Wilson, "The Ants. Berlin: Springer-Verlag, 1990.
96. L. S. Coelho, R. F. T. Neto. Colônia de Formigas: Uma Abordagem Promissora para Aplicações de Atribuição Quadrática e Projeto de Layout. XXIV ENEGEP Florianópolis- SC 2004.
97. R. Beckers, J.L. Deneubourg, S. Goss. Trails and U-turns in the selection of path by the ant *Iasius niger*. *J. theor. Biol.* 159 (1992) 397-415.
98. Dorigo M, Gambardella ML, Middendorf M, Stutzle T. Especial section on "Ant Colony Optimization". *IEEE Trans. Evolut. Comput.* 2002; 6: 317–365.
99. M. Dorigo, T. Stutzle. *Ant colony optimization*. MIT Press, 2004.
100. T. L. A. Carvalho. Aplicação de algoritmos colônia de formigas na reconfiguração e alocação de bancos de capacitores em redes elétricas de distribuição. 2015. Dissertação (Mestrado em Engenharia Elétrica) Curso de Pós-Graduação em Engenharia Elétrica, Universidade Federal da Bahia.
101. R. M. Koide. Algoritmo de colônia de formigas aplicado à otimização de materiais compostos laminados. 2010. Dissertação (Mestrado em Engenharia Mecânica e de Materiais) Curso de Pós-Graduação em Engenharia Mecânica e de Materiais, Universidade Federal do Paraná.

102. M. Dorigo. Aprendizado de otimização e algoritmos naturais (em italiano), Ph.D Thesis Dip. Electronico, Politecnico di Milano, (1992).
103. M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond. Ant colony optimisation: a powerful tool for wavelength selection. *J. Chemometrics* 20 (2007) 146–157.
104. M. Dorigo, T. Stützle. Otimização de colônias de formigas: visão geral e avanços recentes. In: Gendreau M., Potvin JY. (eds) *Handbook of Metaheuristics*. International Series in Operations Research & Management Science, vol. 272. Springer, Cham. 2019.
105. E. Bonabeau, M. Dorigo, G. Theraulaz. *Swarm Intelligence From Natural to Artificial Systems*. Oxford University Press, New York, 1999.
106. E. L. Cesar. Sistemas inteligentes aplicados à coordenação da proteção de sistemas elétricos industriais com relés digitais. 2013. Dissertação (Mestrado em Engenharia Elétrica) Curso de Pós-Graduação em Engenharia Elétrica, Universidade do Estado do Rio de Janeiro.
107. F. Allegrini, A. C. Olivieri. A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis. *Analytica Chimica Acta* 699 (2011) 18– 25.
108. S. Izrailev, D. K. Agrafiotis. Variable selection for QSAR by artificial ant colony systems. *SAR & QSAR Environment. Res.* 13 (2002) 417–423.
109. Q. Shen, J. H. Jiang, J. C. Tao, G.L. Shen, R.Q. Yu, J. Modified ant colony optimization algorithm for variable selection in QSAR modeling: QSAR studies of cyclooxygenase inhibitors. *J. Chem. Inf. Model.* 45 (2005) 1024–1029.
110. P. Ding, Q. D. Su, Q. S. Wu. Ant colony algorithm in chemistry and its application in first derivative fluorescent spectra analyzing. *Chem. J. Chinese Univ.* 23 (2002) 1695–1697.

111. H. Xiaowei, Z. Xiaobo, Z. Jiewen, S. Jiyong, Z. Xiaolei, M. Holmes. Measurement of total anthocyanins content in flowering tea using near infrared spectroscopy combined with ant colony optimization models. *Food Chem.* 164 (2014) 536-543.
112. Y. Zhang, M. Li, L. Qin, Q. & Suk, W. Spectral features extraction for estimation of soil total nitrogen content based on modified ant colony optimization algorithm. *Geoderma*. 333 (2019) 23–34.
113. W.T.S. Vilar, et al. Classification of individual Castor Seeds Using Digital Imaging and Multivariate Analysis. *Journal of the Brazilian Chemical Society*, 2014.
114. M.B.H. Santos, et al. Non-destructive NIR spectrometric cultivar discrimination of castor seeds resulting from breeding programs. *Journal of the Brazilian Chemical Society*, v.25, p. 969-974, 2014.
115. Levine, B.K. 3.19 – validation of Classifiers A2 – Walczak, Steven D. Brown Romá Tauler Beata. IN(Ed). *Comprehensive Chemometrics*. Oxford: Elsevier, 2009. 587-599. ISBN 978-0-444-52701-1.
116. Soares, S.F.C. Um novo método para transferência de modelos de calibração NIR e uma nova estratégia para classificação de sementes de algodão usando imagem hiperespectral NIR. 121f. Tese (Doutorado em Química)- Departamento de Química, Universidade Federal da Paraíba, João Pessoa, 2009.
117. U. T. C. P. Souto. Metodologia baseada em imagem digital, espectros UV-Vis e quimiometria para screening de adulteração de café por cascas e paus. 107f. Tese (Doutorado em Química)- Departamento de Química, Universidade Federal da Paraíba, João Pessoa, 2017.
118. W.Wu, Y. Mallet, B. Walczak, et.al. Comparison of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis, applied to NIR data. *Anal Chim Acta*. 1996;329(3):257-265.
119. S. Goya, R. Gupta. Comparison of genetic algorithm based watermarking techniques using tournament selection approach and roulette wheel approach for fidelity optimization. *Int J Secu Appl*.

- 2015;9(10):405-412.
120. BA. Berg. A brief history of the introduction of generalized ensembles to Markov chain Monte Carlo simulations. *Eur Phys J Special Topics*. 2017;226(4):551-565.
 121. R. Salmerón, CB. García, J. García. Variance inflation factor and condition number in multiple linear regression. *J Stat Comput Sim*. 2018; 88(12):2365-2384.
 122. P. Filzmoose, B. Liebman, K. Varmuz. Repeated double cross validation. *J Chemometr*. 2009;23(4):160-171.
 123. E. Moretto, R. Feet. *Tecnologia de Óleos e Gorduras vegetais na Indústria de Alimentos*. São Paulo: Varela Editora e Livraria, 1998.
 124. EMBRAPA. Definição e Histórico de Canola. Disponível em: <<http://www.cnpt.embrapa.br/culturas/canola/definicao.htm>>. Acessado em: 05/12/2019.
 125. F. F. G. Neto. Classificação de óleos vegetais utilizando voltametria de onda quadrada e métodos quimiométricos. Dissertação de Mestrado em química – Universidade Federal da Paraíba, João Pessoa, 2008. 117p.
 126. F.V. Alves, A. de Sá Junior; D. G. Santana; C. M. dos Santos. Composição química e qualidade fisiológica de sementes de girassol de plantas submetidas à competição intraespecífica. *Ver.Bras.sementes*. 34(3) 2012, 457-465.
 127. J. M. G. Mandarino, .A. C. Roessing, V. T. Benassi. Óleos alimentos funcionais. EMBRAPA 2005.
 128. J. N. Rodrigues, L. A. Gioielli, C. Anton. Propriedades físicas de lipídios estruturados obtidos de misturas de gordura do leite e óleo de milho. *Ciência e Tecnologia de Alimentos*, 23: 226, 2003.
 129. A. Szmczycha-Madeja, M. Welma, P. Pohl. Elemental analysis of teas and their infusions by spectrometric methods. *Trends in Analytical Chemistry*. 35 (2012) 165-181.
 130. N. Paludo. Desenvolvimento e caracterização de kombucha obtida a partir de chá verde e extrato de erva-mate: processo artesanal e

- escala laboratorial. Trabalho de conclusão de curso em engenharia dos alimentos - Dissertação de Mestrado em Ciências dos Alimentos – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2017. 47p.
131. J. D. Lima, P. Mazzafera, W.S. Moraes, R. B. da Silva. Chá: aspectos relacionados à qualidade e perspectivas. *Ciência Rural*, Santa Maria. 39 (2009) 1270-1278.
 132. L.A. FIRMINO, M.P.S. MIRANDA. Polifenóis totais e flavonoides em amostras de chá verde (*Camellia sinensis* L.) de diferentes marcas comercializadas na cidade de Salvador-BA, *Rev. Bras. Pl. Med.* 17 (2015) 436-443.
 133. P.H.G.D. Diniz, M.F. Barbosa, K.D.T.M. Milanez, M.F. Pistonesi, M.C.U. Araújo. Using UV-Vis spectroscopy for simultaneous geographical and varietal classification of tea infusions simulating a home-made tea cup. *Food Chemistry*. 192 (2016) 374-379.
 134. D. Ballabio, V. Consonni. Classification tools in chemistry. Part 1: linear models. PLS-DA. *Anal. Methods*. 16; 5 (2013) 3790–3798.
 135. R. W. Kennard, L. A. Stone. Computer aided design of experiments, *Technometrics*. 11; 1 (1969): 137–148.
 136. Y. Zhang, M. Li, L. Zheng, Q. Qin, W. Suk. Spectral features extraction for estimation of soil total nitrogen content based on modified ant colony optimization algorithm. *Geoderma*. (2019) 333: 23–34.
 137. R. Maesschalck, D. Jouan-Rimbaud, D.L. Massart. The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* 50; 1 (2000): 1–18.
 138. M. D. Scherer. Caracterização óptica do diesel, biodiesel, óleo e suas misturas: Desenvolvimento de uma metodologia alternativa para a quantificação do teor de biodiesel na mistura diesel/biodiesel. Dissertação de Mestrado em Ciência e Tecnologia Ambiental. Universidade Federal da Grande Dourados. Dourados/MS. 2011. P 98.
 139. M. C. Tolentino. Avaliação da qualidade de óleos vegetais sob estresse fotooxidativo e termoxidativo por espectroscopia de uv e rmn de ¹H. Dissertação em Ciência e Tecnologia de Alimentos. Universidade Estadual de Ponta Grossa. 2008. P 99.

140. M. C. Tolentino, *et al.* Avaliação da estabilidade foto-oxidativa dos óleos de canola e de milho em presença de antioxidantes sintéticos. *Ciência Rural* 44 (2014) 728–733.
141. L. S. Ferreira, *et al.* Caracterização de óleos e resinas vegetais da Amazônia por espectroscopia de absorção. *Sci. Plena* 13 (2017) 1–7.
142. R. Perazzini, D. Leonardi, S. Ruggeri, D. Alesiani, A. D'arcangelo, A. Canini. Characterization of *Phaseolus vulgaris* L. Landraces Cultivated in Central Italy. *Plant Foods for Human Nutrition*, 63 (2008) 211.
143. SEAB: Secretaria da Agricultura e do Abastecimento. Disponível em<http://www.agricultura.pr.gov.br/arquivos/File/deral/Prognosticos/2017/Feijao_2016_17.pdf>. Acessado em: 17 jul. 2017.
144. L. P. YOKOYAMA. Aspectos conjunturais da produção de feijão. In: H. AIDAR, J. KLUTHCOUSKI, L. F. STONE (Ed.). Produção de feijoeiro comum em várzeas tropicais. Santo Antônio de Goiás: Embrapa Arroz e Feijão. (2002) 249-292.
145. Companhia Nacional de Abastecimento. Perspectivas para a agropecuária – Safra 2014/2015, Brasília, v. 2 2014.
146. Companhia Nacional de Abastecimento. Acompanhamento as safra brasileira – Safra 2016/2017, Brasília, v. 4 2017.
147. Unifeijão. Feijão do Brasil. Disponível em: <http://unifeijao.com.br/site2013/feijao_brasil.php?txt=1> acessado em: 17 jul. 2017.
148. A. C. C. Moura. Análise físico-químicas e anemáticas antes e após armazenamento em grãos de feijão (*Phaseolus vulgaris* L.) submetidos a diferentes tempos e tipos de secagem. Dissertação de Mestrado em Ciências dos Alimentos – Universidade Federal de Lavras, Lavras, 1998. 70p.
149. M. A. P. Ramalho, A. F. B. Abreu. Como Obter Sucesso na Cultura do Feijoeiro no Estado de Minas Gerais. Circular Técnica 50. Goiás 2002.
150. L. C. Melo *et al.* BRS ESTILO - CULTIVAR DE FEIJÃO CARIOCA COM GRÃOS CLAROS, ARQUITETURA ERETA E ALTO POTENCIAL PRODUTIVO. Embrapa Arroz e Feijão. 2011.

151. O. F. da Silva, A. E. Wander. Impacto Econômico da Cultivar de Feijão Tipo Carioca BRS Pontal. Comunicado Técnico 172 ISSN 1678-961X. EMBRAPA, Santo Antônio de Goiás, GO. 2009.
152. T. L. P. O. de Souza et. al. BRS FC401 RMD: Cultivar de Feijão Carioca Geneticamente Modificada com Resistência ao Mosaico-dourado. Comunicado Técnico 235 ISSN 1678-961X. EMBRAPA. Santo Antônio de Goiás. 2016
153. M.S. Kumar, S.S. Chakravarthy, P.R. Babu, K.V. Rao, V.D. Reddy. Classification of cytochrome P450s in common bean (*Phaseolus vulgaris* L.). *Plant Systematics and Evolution* 301 (2015) 211-6.
154. J. Schmutz, P.E. Mcclean, S. Mamidi, et al. A reference genome for common bean and genome-wide analysis of dual domestications. *Nat Genet* 46 (2014) 707-13.
155. Agriculture USDOHaHSaUSDO, 2015-2020. Dietary Guidelines for Americans.
156. I. Hayat, A. Ahmad, T. Masud, A. Ahmed, S. Bashir. Nutritional and health perspectives of beans (*Phaseolus vulgaris* L.): an overview. *Crit Rev Food Sci Nutr* 54 (2014) 580-92.
157. V. C. SGARBIERI. Estudo do conteúdo e de algumas características das proteínas e sementes de plantas leguminosas. *Ciência e Cultura*, São Paulo, 32(1980) 78-84.
158. F. Goycoolea, E. González, J. M. Barron. (*Phaseolus vulgaris*) sobre el contenido de taninos y valor nutritivo de las proteínas. *Archivos Latino-Americanos de Nutrición*, Guatemala, 15 (1990) 263-273.
159. P.B. Geil, J.W. Anderson. Nutrition and health in implications of dry beans: A review. *Journal of the American College of Nutrition*, 13 (1994) 549.
160. B. B. G. Tartarin, W. S. Barros, L. C. D. de Souza. Principais pragas na cultura do feijoeiro comum. Ver. *Conexão Eletrônica - Três Lagoas*, MG. 13 (2016).

161. M. C. SILVA-FILHO, M. C. FALCO. Plantas transgênicas no melhoramento. Recursos genéticos e melhoramento plantas. Rondonópolis: Fundação MT, 1011p, 2001.
162. R. Choze. Estudo de feijão transgênicos e convencionais através da RMN HR-MAS. 2011. Tese (Doutorado em Química) - Curso de Pós-Graduação em Química, Universidade Federal da Goiás.
163. B. Fleming. Mosaico dourado: 6 dicas para combater a praga do feijão. Disponível em: < <http://sfagro.uol.com.br/mosaico-dourado-6-dicas-para-combater-praga-do-feijao/> > acessado em: 17 jul. 2017.
164. L. Amorim, J. A. M. Rezende, A. Bergamin Filho, L. E. A. Camargo. Doenças de Plantas Cultivadas. Manual de Fitopatologia. Editora Agronomia Ceres Ltda, Ouro Preto-MG. 2016.
165. EMBRAPA.. Um sistema e um feijão transgênico para enfrentar o mosaico-dourado. Disponível em: < <https://www.embrapa.br/en/busca-de-noticias/-/noticia/8534579/um-sistema-e-um--feijao-transgenico-para-enfrentar-o-mosaico-dourado> > Acessado em: 17 jul. 2017.
166. G. Castillo. O mosaico dourado no feijoeiro. Disponível em: < <https://3rlab.wordpress.com/2016/08/19/o-mosaico-dourado-no-feijoeiro/> > ACESSADO EM: 17 JUL. 2017.
167. J. Rebolla. O feijão resistente ao vírus do mosaico dourado. Disponível em: < <http://jornalgggn.com.br/blog/luisnassif/o-feijao-resistente-ao-virus-do-mosaico-dourado> > Acessado em: 17 jul. 2017.
168. Entenda os transgênicos: alimentos, animais e microrganismos podem se encaixar nesse grupo. Disponível em: < <http://www.ecycle.com.br/component/content/article/62/2384-organismos-transgenicos-o-que-sao-dna-geneticamente-modificados-ogm-milho-soja-algodao-biosseguranca-monsanto-greepeace-idec-consequencias-saudeo-humana-alergia-cancer-biodiversidade-superpragas-trangenias-agrossistemas-como-evitar-alternativas-organicos.html> > Acessado em: 20 jul. 2017.
169. H.A. PERR. Children and genetically engineered food: potentials and problems. *Journal of Pediatric Gastroenterology and Nutrition*, 35(2002) 475, 2002.

170. TAYLOR, S.; HEFLE, S.L. Will genetically modified foods be allergenic. *Journal of Allergy and Clinical Immunology*, 107 (2001) 765.
171. ASTWOOD, J.D.; BANNON, G.A; DOBERT, R.L.; FUCHS, R.L. Food biotechnology and genetic engineering. In: METCALFE, D.D.; SAMPSON, H.A.; SIMON, R.A.9eds). *Food Allergy: Adverse Reactions to Food and Food Additives*, **3**: 38, 2003.
172. NEWELL, C.A .PLANT TRANSFORMATION TECHNOLOGY. Developments and applications. *Molecular Biotechnology*, 16 (2000) 53.
173. Eco. O que são alimentos transgênicos. Disponível em: < <http://www.oeco.org.br/dicionario-ambiental/27355-o-que-sao-alimentos-transgenicos/>> Acessado em: 17 jul. 2017.
174. A. Floriani. Mercado de transgenicos movimentava bilhões de dólares. Disponível em: < <http://www.terra.com.br/reporterterra/transgenicos/mundo.htm>> Acessado em: 17 jul. 2017.
175. A. Germini, A. Zanetti, C. Salati, S. Rossi, C. Forré, S. Schmid, R. Marchelli. Development of a seven-target Multiplex PCR for the simultaneous detection of transgenic soybean and maize in feeds and foods. *J Agric Food Chem* 52 (2004) 3275-3280.
176. P. Cardarelli, M. R. Branquinho, R.T.B. Ferreira, F.P. Cruz, A. L. Gemal. Detection of GMO in food products in Brazil: the INCQS experience. *Food Control* 16 (2005) 859-866.
177. ISAAA: Serviço Internacional para a Aquisição de Aplicações de Agrobiotecnologia). Disponível em: <<https://isaaa.org/resources/publications/briefs/49/executivesummary/pdf/B49-ExecSum-Portuguese-Brazil.pdf>> acessado em 22/07/2017
178. ISAAA: Resumo ISAAA. Disponível em:< <http://www.pioneersementes.com.br/biotecnologia/resumo-isaaa>> Acessado em: 18 jul.2017.
179. Faria JC, Valdisser PaMR, Nogueira EOPL, Aragão FJL, 2014. RNAi-based Bean golden mosaic virus-resistant common bean (Embrapa

- 5.1) shows simple inheritance for both transgene and disease resistance. *Plant Breeding* 133, 649-53.
180. BRASIL. Ministério da Agricultura e Reforma Agrária. Regras para análise de sementes. Brasília: SNDA/DNDV/CLAV, 2009. 365p.
 181. C. S. M. Cunha, M. Â. A. Tillmann, F. A. Villela, L. B. Dode, F. Balerini. Comparação de métodos na detecção de sementes de soja geneticamente modificada resistente ao glifosato. *Revista Brasileira de Sementes*, vol. 27 (2005) 167-175.
 182. F. R. Conceição, A. N. Moreira, P. C. Binsfeld. Detecção e quantificação de organismos geneticamente modificados em alimentos e ingredientes alimentares. *Ciência Rural*, 36 2006.
 183. E. Micheline, P. Simoni, L. Cevenini, L. Mezzanotte, A. Roda. New trends in bioanalytical tools for the detection of genetically modified organisms: an update . *Anal Bioanal Chem.* 392 (2008) 355–367.
 184. A. K. Deisingh, N. Badrie. DETECTION APPROACHES FOR GENETICALLY MODIFIED ORGANISMS IN FOODS. *Food Res Int* 38 (2005)639–649.
 185. K. S. Kumar, S. H. Kang. Ultra-fast simultaneous analysis of genetically modified organisms in maize by microchip electrophoresis with LIF detector. *Electrophoresis* 28 (2007) 4247–425.
 186. J.M. Rodríguez-Nogales, A. Cifuentes, M.C. García, M.L. Marina. Characterization of protein fractions from Bt-transgenic and non-transgenic maize varieties using perfusion and monolithic rpc. Maize differentiation by multivariate analysis. *J Agric Food Chem* 55 (2007)3835–3842
 187. L. Xie, Y. Ying, T. J. Ying. Quantification of Chlorophyll Content and Classification of Nontransgenic and Transgenic Tomato Leaves Using Visible/ Near-Infrared Diffuse Reflectance Spectroscopy. *Agric. Food Chem.* 55 (2007) 4645-4650.
 188. S. Zhu, J. Liang, L. Yan. Study on Rapid Identification methods of Transgenic Rapeseed Oil Based on Near Infrared Spectroscopy.

189. R. Rukun, L. Yunbo, H. Kunlun, et al.: Application of near-infrared diffuse reflectance spectroscopy to the detection and identification of transgenic corn .J. Spectroscopy and Spectral Analysis. 25 (2005) 1581-1583.
190. L.F.S. PEREIRA. Espectroscopia no *NIR* e processamento de imagens digitais para classificação de mamão papaia (*carica papaya l.*) *golden*. Dissertação (Mestrado)- Universidade Estadual de Campinas- UNICAMP, 80p.2017.

Received: 20 January 2020 | Revised: 8 June 2020 | Accepted: 31 July 2020
DOI: 10.1002/cem.3292

SPECIAL ISSUE - RESEARCH ARTICLE

JOURNAL OF
CHEMOMETRICS WILEY

Ant colony optimization for variable selection in discriminant linear analysis

Aline S. Pontes¹ | Alisson Araújo² | Weverton Marinho² |
Paulo H. Gonçalves Dias Diniz³ | Adriano de Araújo Gomes⁴ |
Hector C. Goicoechea⁵ | Edvan C. Silva¹ | Mario C.U. Araújo¹

¹Departamento de Química, Laboratório de Automação e Instrumentação em Química Analítica e Quimiometria (LAQA) Universidade Federal da Paraíba, CCEN, João Pessoa, Paraíba, Brazil

²Faculdade de Química, Instituto de Ciências Exatas da Universidade Federal do Sul e Sudoeste do Pará, Marabá, Pará, Brazil

³Programa de Pós-Graduação em Química Pura e Aplicada, Universidade Federal do Oeste da Bahia, R. Prof. José Seabra de Lemos 31, Barreiras, Bahia, 47808-021, Brazil

⁴Laboratório de Quimiometria e Instrumentação Analítica-LQIA, Instituto de Química da Universidade Federal do Rio Grande do Sul, Porto Alegre, Rio Grande do Sul, Brazil

⁵Laboratorio de Desarrollo Analítico y Quimiometria (LADAQ), Cátedra de Química Analítica I, Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral-CONICET, Santa Fe, Santa Fe, Argentina

Correspondence

Aline S. Pontes and Edvan C. Silva,
Departamento de Química, Laboratório de Automação e Instrumentação em Química Analítica e Quimiometria (LAQA)
Universidade Federal da Paraíba, CCEN
Caixa Postal 5093, CEP 58051-970 João Pessoa, Brazil.
Email: alinepontes85@hotmail.com;
edvan@quimica.ufpb.br

Adriano de Araújo Gomes, Programa de Pós-Graduação em Química Pura e Aplicada, Universidade Federal do Oeste da Bahia, Campus Reitor Edgard Santos
Rua Bertogã, 892, Bairro Morada Nobre I,

Abstract

A new algorithm using ant colony optimization (ACO) for selection of variables in linear discriminant analysis (LDA) is presented. The role of ACO is explored in the context of LDA classification in which spectral variable multicollinearity is a known cause of generalization problems. The proposed ACO-LDA presents a metaheuristic that mimics the ant's cooperative behavior, randomly depositing pheromones at vector elements corresponding to the most relevant variables. Such cooperative ant-like behavior, which is absent in the genetic algorithm, increases the probability of discarding noninformative variables, favoring construction of more parsimonious models than genetic algorithm-linear discriminate analysis (GA-LDA). The classification performance of ACO-LDA is assessed in two case studies: (i) classification of edible vegetable oils (with respect to base oil) via ultraviolet-visible (UV-Vis) spectrometry and (ii) simultaneous classification of tea samples with respect to type and geographic origin via near-infrared (NIR) spectrometry. In the first study, ACO-LDA was tested in a data set involving wide absorption bands in the UV region with low-resolution and strong spectral overlapping. In the second study, its capacity to manage a data matrix with high dimensionality was evaluated. In both studies, ACO-LDA selected a small subset of variables, which led to correct classifications for almost all of the samples, achieving a performance level similar to the well-established partial least squares-discriminant analysis (PLS-DA), and considerably better than GA-LDA. The use of ACO to select LDA classification variables can minimize generalization problems commonly associated with multicollinearity.

KEYWORDS

ant colony optimization, edible oil and tea, linear discriminant analysis, near-infrared and ultraviolet-visible spectrometry, variable selection