

RESSALVA

Atendendo solicitação do(a) autor(a), o texto completo será disponibilizado somente a partir de 17/08/2019.

UNESP
Faculdade de Engenharia do Campus de Guaratinguetá

Guaratinguetá
2017

LEANDRO VALIM DE FREITAS

NOVAS ESTRATÉGIAS NA MODELAGEM MULTIVARIADA DE
PARÂMETROS DE QUALIDADE NUM PROCESSO QUÍMICO INDUSTRIAL

Tese apresentada à Faculdade de
Engenharia do Campus de
Guaratinguetá, Universidade Estadual
Paulista, para a obtenção do título de
Doutor em Engenharia Mecânica na
área de Gestão e Otimização.

Orientador: Prof. Dr. Messias Borges Silva
Co-orientador: Prof. Dr. Fernando Augusto Silva Marins

Guaratinguetá
Agosto 2017

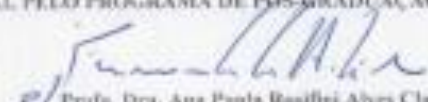
F866n	Freitas, Leandro Valim de Novas estratégias na modelagem multivariada de parâmetros de qualidade num processo químico industrial / Leandro Valim de Freitas – Guaratinguetá, 2017 160 f. : il. Bibliografia: f. 92-103 Tese (doutorado) – Universidade Estadual Paulista, Faculdade de Engenharia de Guaratinguetá, 2017. Orientador: Prof Dr Messias Borges Silva Coorientador: Prof Dr Fernando Augusto Silva Marins 1. Taguchi, Metodo de (Controle de qualidade). 2. Analise multivariada. 3. Controle de processo . I. Título CDU 658.56(043)
-------	---

LEANDRO VALIM DE FREITAS

ESTA TESE FOI JULGADA ADEQUADA PARA A OBTENÇÃO DO TÍTULO DE
"DOUTOR EM ENGENHARIA MECÂNICA"

PROGRAMA: ENGENHARIA MECÂNICA
ÁREA: GESTÃO E OTIMIZAÇÃO

APROVADA EM SUA FORMA FINAL PELO PROGRAMA DE PÓS-GRADUAÇÃO


Prof. Dra. Ana Paula Rosilini Alves Claro
Coordenadora

BANCA EXAMINADORA:


Prof. Dr. MESSIAS BORGES SILVA
Orientador/UNESP-FEG


Prof. Dr. MARCELA APARECIDA GUERREIRO
MACHADO DE FREITAS
UNESP-FEG


Prof. Dr. ANEIRSON FRANCISCO DA SILVA
UNESP-FEG


Prof. Dr. FABRICIO MACIEL GOMES
EEL-USP


Prof. Dr. ADRIANO FRANCISCO SIQUEIRA
EEL-USP

Agosto de 2017

AGRADECIMENTOS

- À Deus, que me concedeu saúde e força de vontade por estar aqui.

- Expresso também minha gratidão às pessoas que me incentivaram no desenvolvimento da pesquisa.

- A minha querida família: esposa Ana Paula e filha Maria Clara que são o meu alicerce.

- Aos meus pais que sempre me lembram da importância da paciência e da humildade.

- Agradeço a todos os professores que já tive na vida e, em especial, os orientadores e amigos Prof. Dr. Messias Borges Silva e Prof. Dr. Fernando Augusto Silva Marins do Departamento de Produção da FEG-UNESP pela paciência, boa vontade em ajudar e por estarem sempre me motivando durante nossas conversas. Agradeço sempre também as oportunidades que me ofereceram na área acadêmica.

FREITAS, L. V. **Novas estratégias na modelagem multivariada de parâmetros de qualidade num processo químico industrial**. 2017. 80 f. Tese (Doutorado em Engenharia Mecânica) – Faculdade de Engenharia do Campus de Guaratinguetá, Universidade Estadual Paulista, Guaratinguetá, 2017.

RESUMO

Este trabalho tem por objetivo desenvolver e validar modelos matemáticos multivariados de predição com a finalidade de monitorar num processo químico de manufatura. Os produtos derivados têm especificações baseadas em propriedades físico-químicas que podem variar de maneira significativa com a sua modificação do elenco de matérias-primas, mantendo-se as mesmas condições de controle de produção, o que compromete os padrões de qualidade dos seus produtos finais. Isso leva à necessidade de determiná-las ou prevê-las com a maior frequência possível em detrimento às análises pontuais tradicionais em laboratório. Foi possível modelar matematicamente as propriedades pela técnica multivariada de Máquina de Vetores Suporte (SVM – *Support Vector Machines*) a partir de dados experimentais e compará-la com a abordagem multivariada tradicional por Mínimos Quadrados Parciais (PLS – *Partial Least Squares*). Como o algoritmo PLS, amplamente utilizado no tratamento de dados desta natureza não possui o parâmetro de regularização como as máquinas, foi proposto o algoritmo PLS–Taguchi na tentativa de auxiliar a sua configuração e otimizar seus parâmetros para contornar essa ausência.

PALAVRAS-CHAVE: Análise Multivariada, Taguchi, Monitoramento de Processos

FREITAS, L. V. **New strategies in the multivariate modeling of quality parameters in an industrial chemical process.** 2017. 80 l. Thesis (Ph.D. in Mechanical Engineering) - Faculty of Engineering of Guaratinguetá Campus, State University of São Paulo, Guaratinguetá, 2017.

ABSTRACT

This work aims to develop and validate multivariate predictive mathematical models for the purpose of monitoring in a chemical manufacturing process. Derivatives have specifications based on physicochemical properties that can vary significantly with their modification of the cast of raw materials, while maintaining the same conditions of production control, which compromises the quality standards of their final products . This leads to the need to determine or predict them as often as possible over traditional laboratory analyzes. It was possible to mathematically model properties using the multivariate Vector Supporting Machine (SVM) technique from experimental data and compare it with the traditional multivariate approach by Partial Least Squares (PLS). As the PLS algorithm, widely used in the treatment of data of this nature does not have the regularization parameter as the machines, the PLS-Taguchi algorithm was proposed in an attempt to aid its configuration and optimize its parameters to overcome this absence.

KEYWORDS: Multivariate Analysis, Taguchi, Process Monitoring

LISTA DE FIGURAS

Figura 1 – Publicações com relação à palavra-chave SVM.....	18
Figura 2 – Publicações com relação às palavras-chave SVM x produto.....	19
Figura 3 – Etapas da Pesquisa.....	21
Figura 4 – Esquema da tecnologia para aquisição do banco de dados.....	29
Figura 5 – Conjunto de espectros de absorção da amostra 1.....	30
Figura 6 – Espaço bidimensional de duas classes de amostras.....	38
Figura 7 – Espaço tridimensional de duas classes de amostras.....	38
Figura 8 – Hiperplano ótimo de separação das duas classes de amostras.....	39
Figura 9 – Hiperplano de separação das duas classes de amostras.....	40
Figura 10 – Hiperplano de separação das duas classes de amostras.....	40
Figura 11 – Hiperplano de separação das duas classes de amostras.....	41
Figura 12 – Hiperplano ótimo de separação das duas classes de amostras.....	41
Figura 13 – Pontos de mínimo.....	43
Figura 14 – Espectros das 197 amostras levantadas na pesquisa.....	48
Figura 15 – Amplitude dos coeficientes de correlação.....	49
Figura 16 – Amplitude dos coeficientes de autocorrelação das amostras estudadas.....	50
Figura 17 – Escores dos dois primeiros componentes principais de uma PCA das..... amostras estudadas.....	51
Figura 18 – Histograma para a propriedade P1 do ponto inicial.....	51
Figura 19 – Histograma para a propriedade P2 de 10% recuperados.....	52
Figura 20 – Histograma para a propriedade P3 de 50% recuperados.....	52
Figura 21 – Histograma para a propriedade P4 de 90% recuperados.....	53
Figura 22 – Histograma para a propriedade P5 ponto final de ebulição (PFE).....	53
Figura 23 – Histograma para a propriedade P6.....	54
Figura 24 – Histograma para a propriedade P7.....	54
Figura 25 – Histograma para a propriedade P8.....	55
Figura 26 – Histograma para a propriedade P9 a -20°C.....	55
Figura 27 – Gráfico dos efeitos dos fatores sobre a média para P1 com relação à Y1....	60
Figura 28 – Gráfico de contorno para P1 com relação à Y ₁	65
Figura 29 – Gráfico de superfície para P1 com relação à Y ₁	66
Figura 30 – Gráfico de contorno para P1 com relação à Y ₂	66

Figura 31 – Gráfico de superfície para P1 com relação à Y_2	67
Figura 32 – Gráfico de contorno para P1 com relação à Y_3	67
Figura 33 – Gráfico de superfície para P1 com relação à Y_3	68
Figura 34 – Gráfico de contorno para P1 com relação à Y_4	68
Figura 35 – Gráfico de superfície para P1 com relação à Y_4	69
Figura 36 – Escores das duas primeiras variáveis latentes para P1.....	69
Figura 37 – Valores medidos versus valores preditos para P1.....	70
Figura 38 – Variância explicada versus número de variáveis latentes para P1.....	70
Figura 39 - Valores preditos versus valores medidos para P1 Kernel RBF.....	73
Figura 40 – Valores preditos versus valores medidos para P2 10% Kernel RBF	74
Figura 42 – Valores preditos versus valores medidos para P3 50% Kernel Sigmoidal	74
Figura 43 – Valores preditos versus valores medidos para P4 90% Kernel Linear.....	75
Figura 44 – Valores preditos versus valores medidos para P5 Kernel Linear.....	75
Figura 45 – Valores preditos versus valores medidos para P6 Kernel Linear.....	76
Figura 46 – Valores preditos versus valores medidos para P7 Kernel Linear.....	76
Figura 47 – Valores preditos versus valores medidos para P8 Kernel RBF.....	77
Figura 48 – Valores preditos versus valores medidos para P9 Kernel RBF.....	78
Figura 49 – Teste F para os conjuntos RMSECV e RMSEP: modelo P1 via PLS.....	78
Figura 50 – Teste F para os conjuntos RMSECV e RMSEP: modelo para P2 10% via PLS.....	79
Figura 51 – Teste F para os conjuntos RMSECV e RMSEP: modelo para P3 50% via PLS.....	79
Figura 52 – Teste F para os conjuntos RMSECV e RMSEP: modelo para P4 90% via PLS.....	79
Figura 53 – Teste F para os conjuntos RMSECV e RMSEP: modelo para P5 via PLS.....	80
Figura 53 – Teste F para os conjuntos RMSECV e RMSEP: P6 via. PLS.....	80
Figura 54 – Teste F para os conjuntos RMSECV e RMSEP: P7 via PLS.....	81
Figura 55 – Teste F para os conjuntos RMSECV e RMSEP: P8 via PLS.....	81
Figura 56 – Teste F para os conjuntos RMSECV e RMSEP: P9 via PLS.....	82
Figura 57 – Teste F para os conjuntos RMSECV e RMSEP: P1 via SVM RBF.....	82

Figura 58 – Teste F para os conjuntos RMSECV e RMSEP: P2 10% via SVM RBF	83
Figura 59 – Teste F para os conjuntos RMSECV e RMSEP: P3 50% SVM Sigmoidal	83
Figura 60 – Teste F para os conjuntos RMSECV e RMSEP: P4 90% SVM Linear.....	84
Figura 61 – Teste F para os conjuntos RMSECV e RMSEP: P5 SVM Linear.....	84
Figura 62 – Teste F para os conjuntos RMSECV e RMSEP: P6 SVM Linear.....	85
Figura 63 – Teste F para os conjuntos RMSECV e RMSEP: P7 SVM Linear.....	85
Figura 64 – Teste F para os conjuntos RMSECV e RMSEP: P8 SVM Linear.....	86
Figura 64 – Teste F para os conjuntos RMSECV e RMSEP: P9 SVM RBF.....	96

LISTA DE TABELAS

Tabela 1 – Ocorrências das palavras-chave nas bases de dados Scopus e Web of Science.....	19
Tabela 2 – ANOVA para arranjos ortogonais de Taguchi.....	46
Tabela 3 – Níveis dos Fatores para o PLS-Taguchi.....	56
Tabela 4 – Variáveis respostas e objetivos de otimização.....	56
Tabela 5 – AO L ₉ para destilação P1.....	57
Tabela 6 – AO L ₉ para P2 10% recuperados.....	57
Tabela 7 – AO L ₉ para P3 50% recuperados.....	57
Tabela 8 – AO L ₉ para P4 90% recuperados.....	58
Tabela 9 – AO L ₉ para P5.....	58
Tabela 10 – AO L ₉ para P6.....	58
Tabela 11 – AO L ₉ para P7.....	59
Tabela 12 – AO L ₉ para P8.....	59
Tabela 13 – AO L ₉ para P9 a -20°C.....	59
Tabela 14 – ANOVA para P1.....	61
Tabela 15 – ANOVA para P2 10% recuperados.....	61
Tabela 16 – ANOVA para P3 50% recuperados.....	61
Tabela 17 – ANOVA para P4 90% recuperados.....	61
Tabela 18 – ANOVA para P5.....	61
Tabela 19 – ANOVA para P6.....	62
Tabela 20 – ANOVA para P7.....	62
Tabela 21 – ANOVA para P8.....	62
Tabela 22 – ANOVA para P9 a -20°C.....	62
Tabela 23 – Relatório <i>Desirability</i> para P1.....	63
Tabela 24 – Relatório <i>Desirability</i> para P2 10% recuperados.....	63
Tabela 25 – Relatório <i>Desirability</i> para P3 50% recuperados.....	63
Tabela 26 – Relatório <i>Desirability</i> para P4 90% recuperados.....	63
Tabela 27 – Relatório <i>Desirability</i> para P5.....	63
Tabela 28 – Relatório <i>Desirability</i> para P6.....	64
Tabela 29 – Relatório <i>Desirability</i> para P7.....	64

Tabela 30 – Relatório <i>Desirability</i> para P8.....	64
Tabela 31 – Relatório <i>Desirability</i> para P9 a -20°C.....	64
Tabela 32 – Parâmetros de modelagem via SVM.....	71

LISTA DE ABREVIATURAS E SIGLAS

AHP	Analytic Hierarchy Porcess
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
API	American Petroleum Institute
ASTM	American Standard for Testing and Materials
DEA	Data Envelopment Analysis
DOE	Design of Experiments
DMAIC	Define, Measure, Analyse, Improve and Control
ELECTRE	Elimination and Choice Expressing Reality
FT	Fourier Transform
GAIA	Geometrical Analysis for Iterative Assistance
I-MR	Individual – Moving Range
JFTOT	Jet Fuel Thermal Oxidation Teste
KDD	Knowledge Discovery in Databases
LS-SVM	Least Squares Support Vector Machines
MCDM	Multiple Criteria Decision Making
MC-RPLS	Monte Carlo Sampling Recursive Partial Least Squares
MLR	Multiple Linear Regression
PCA	Principal Component Analysis
PCR	Principal Components Analysis
PLS	Partial Least Squares
PLS-DA	Partial Least Squares Discriminant Analysis
PROMETHEE	Preference Ranking Organization Method for Enrichment Evaluations
RBF	Radial Basis Function
RMSECV	Root Mean Square Error of Cross Validation
RMSEP	Root Mean Square Error of Prediction
RSM	Response Surface Method
SIMCA	Soft Independent Modeling of Class Analogy
SPA	Successive Projections Algorithm
SVM	Support Vector Machines

SUMÁRIO

1	INTRODUÇÃO.....	14
1.1	Contextualização do Problema.....	14
1.2	Objetivos.....	15
1.3	Justificativas e Contribuições da Tese.....	15
1.4	Delimitações do trabalho.....	20
1.5	Método e Classificação da Pesquisa.....	20
1.6	Estrutura do Trabalho.....	24
2	REFERENCIAL TEÓRICO.....	25
2.1	Aplicações de métodos multivariados de <i>Big Data Mining</i>	25
2.2	Revisão sobre os métodos utilizados.....	28
2.2.1	Ciência e Tecnologia envolvidas na mineração dos dados.....	28
2.2.2	Métodos multivariados de <i>Big Data Mining</i>	30
2.2.2.1	Redução de Dimensionalidade.....	30
2.2.2.2	Aumento de Dimensionalidade.....	37
2.2.2.3	Planejamento de Experimentos.....	44
2.2.2.4	Experimentos de Taguchi.....	45
2.2.2.4.1	Otimização de Experimentos pela função <i>Desirability</i>	46
3	RESULTADOS.....	48
3.1	Fase Definir.....	48
3.2	Fase Medir.....	49
3.3	Fase Analisar.....	49
3.3.1	Questão a Autocorrelação e Estratégia de Amostragem.....	49
3.4	Fase Melhorar.....	56
3.4.1	Proposta do Algoritmo PLS-Taguchi.....	56
3.4.2	Proposta do Algoritmo LS-SVM.....	71
3.5	Fase Controlar.....	79
4	CONCLUSÕES.....	87
4.1	Verificação dos objetivos e respostas às questões de pesquisa.....	87
4.2	Sugestões para continuidade do trabalho.....	87
5	REFERÊNCIAS.....	89
	ANEXOS.....	101

1 INTRODUÇÃO

1.1 Contextualização do Problema

O petróleo possui em sua composição química uma mistura heterogênea e extremamente complexa de compostos orgânicos e inorgânicos, que sofrem variações ao longo do tempo, mesmo para amostras retiradas de um mesmo poço. Isto leva a necessidade de monitorar nos processos de refino os parâmetros de qualidade desta matéria-prima e de seus derivados, que são refletidos por meio de suas propriedades físico-químicas, com a maior frequência possível em detrimento das análises pontuais tradicionais em laboratórios petroquímicos (NICODEM et al., 1997).

Um dos principais derivados do petróleo e principal combustível para aeronaves é o querosene de aviação, apesar de sua importância na matriz energética de um país, são escassos os trabalhos relacionados ao seu sistema produtivo. Estudos recentes envolvendo petróleo e derivados mostraram a potencialidade de avançadas técnicas analíticas instrumentais rápidas para estimar propriedades físico-químicas e termodinâmicas (FILGUEIRAS et. al, 2015; DUARTE et al., 2016). Nestes métodos, a conversão desta resposta instrumental na informação de interesse, requer a utilização de técnicas multivariadas de *Data Mining* (mineração de dados) (BALABIN, SAFIEVA, LOMAKINA, 2010) e *Data Mining* em grande escala – *Big Data* (KRANJC et al., 2017).

Segundo Amaral (2001), a mineração de dados é a busca por padrões úteis em conjunto de dados e tem recebido outros diversos nomes como prospecção ou descoberta de conhecimento em banco de dados, (*Knowledge Discovery in Databases, KDD*), extração de conhecimento, descoberta de informação, dentre outros. É definida como o processo automático, ou semi-automático de extração de conhecimento potencialmente útil e de identificação de padrões, tendências, associações e dependências, previamente desconhecidos, de bases de dados (OLAFSSON et al., 2008). O uso destas técnicas, que se baseiam fortemente em métodos estatísticos, inteligência artificial, aprendizagem de máquina (HARDING et al., 2006), em conjunto com métodos da Pesquisa Operacional (LAU et al., 2015), auxilia na resolução destes problemas interdisciplinares (PAN; DIAS, 2017).

1.2 Objetivos

Este trabalho tem por objetivo geral desenvolver modelos matemáticos multivariados de previsão via máquina de vetores suporte por mínimos quadrados (*LS-SVM – Least Squares Support Vector Machines*) com vistas ao monitoramento da qualidade no processamento de produtos em uma indústria química.

Como objetivos específicos, buscou-se:

- Comparar estatisticamente o desempenho dos modelos de previsão por *SVM* com a abordagem tradicional por mínimos quadrados parciais (*PLS*).
- Testar os modelos desenvolvidos num processo real industrial mediante validação estatística.
- Integrar Projeto de Experimentos com Análise Multivariada na otimização dos modelos de *PLS* propostos.

1.3 Justificativas e Contribuições da Tese

Sinais de radiação infravermelha associados aos métodos multivariados têm sido amplamente utilizados na indústria química no monitoramento de processos e previsão da qualidade de produtos acabados. Vários trabalhos foram escritos sobre aplicações da Análise Multivariada combinada à espectroscopia infravermelha na predição de algumas propriedades da indústria química (HONORATO et al., 2008; KARDAMAKIS; PASADAKIS, 2010; FREITAS et al., 2013; SILVA et al., 2014; BRAGA; JUNIOR; MARTINS, 2014; TOMREN; BARTH, 2014). Os métodos mais popularmente empregados ao longo dos anos foram Regressão Linear Múltipla (*MLR – Multiple Linear Regression*) (GALVÃO et al., 2006), Regressão por Componentes Principais (*PCR – Principal Components Regression*) (PASTI et al., 2001; PRAKASH; MISHRA, 2013) e Regressão por Mínimos Quadrados Parciais (*PLS – Partial Least Squares*) (KHANMOHAMMADI; GARMARUDI; GUARDIA, 2012).

Os sinais espectrais são as variáveis de entrada que serão relacionadas com propriedades de interesse e, são correlacionados entre si, portanto, são ditos colineares ou multicolineares (NAES et al., 2002). Ficou comprovado que o método *MLR* apresenta sérias limitações devido a essa questão (MUTIHAC; MUTIHAC, 2008). No entanto, para minimizar essas dificuldades, a seleção de variáveis pode anteceder a *MLR*, mediante conhecimento prévio de quais variáveis

se relacionam com a propriedade de interesse (não se aplica a este trabalho), ou ainda, por meio de métodos como o Algoritmo das Projeções Sucessivas (*SPA – Successive Projections Algorithm*) (GALVÃO et al., 2008; OTERO et al., 2011; SOARES et al., 2013) e metaheurísticas (PESSOA, et al., 2015).

Segundo Naes et al. (2002), neste ambiente de muitas variáveis e colineares, a alternativa mais adequada à regressão clássica é a abordagem por componentes principais como PCR e PLS, uma vez que, os componentes principais são ortogonais entre si, então, o problema de multicolinearidade é contornado (NAES; MARTENS, 1984). Juntamente com suas variantes e modificações, o PLS dentre estes vem se mostrando ser o mais adequado (ANDRADE et al., 1997; GONZAGA; PASQUINI, 2010; LIRA et al., 2010; KILLNER; ROHWEDDER; PASQUINI, 2011; ZHANG et al., 2012; BASSBASI et al., 2013), pois diferentemente do PCR cada variável latente é obtida de forma a maximizar a covariância entre Y e todas as possíveis combinações lineares das variáveis X (JOLLIFFE, 2002).

Ao longo dos últimos anos vários pesquisadores têm se preocupado com uma grande questão no método PLS: assumir que a relação espectro – propriedade é linear, no entanto, essa premissa nem sempre é válida para problemas industriais que carregam não linearidades intrínsecas como no caso de produtos complexos do petróleo (FALLA et al., 2006; BALABIN; SAFIEVA, 2008; FILGUEIRAS; ALVES; POPPI, 2014).

Uma proposta bastante disseminada na literatura para contornar essa questão, é o uso das Redes Neurais Artificiais (ANN – *Artificial Neural Networks*) (COELHO et al., 1999; PASQUINI; BUENO, 2007; BALABIN; LOMAKINA; SAFIEVA, 2011; MONEDERO et al., 2012; MENG; JIA; WANG, 2014). Todavia, pesquisadores pontuaram sérias questões que podem comprometer a aplicação de redes neurais em dados espectrais (XU et al., 2017) para produtos de petróleo (THISSEN et al., 2004; BALABIN; SAFIEVA; LOMAKINA, 2007; HERNÁNDEZ et al., 2009):

- Natureza estocástica dos dados, gerando a necessidade de repetir severas vezes o conjunto de treinamento;
- Múltiplas soluções finais com ótimos locais;
- Dependência do resultado final com os parâmetros iniciais;
- Forte tendência de sobreajuste dos dados de treinamento devido ausência de parâmetro de regularização do risco empírico e estrutural;
- Tempo computacional elevado: conjunto de treinamento pode levar, horas, às vezes dias dependendo da configuração dos computadores.

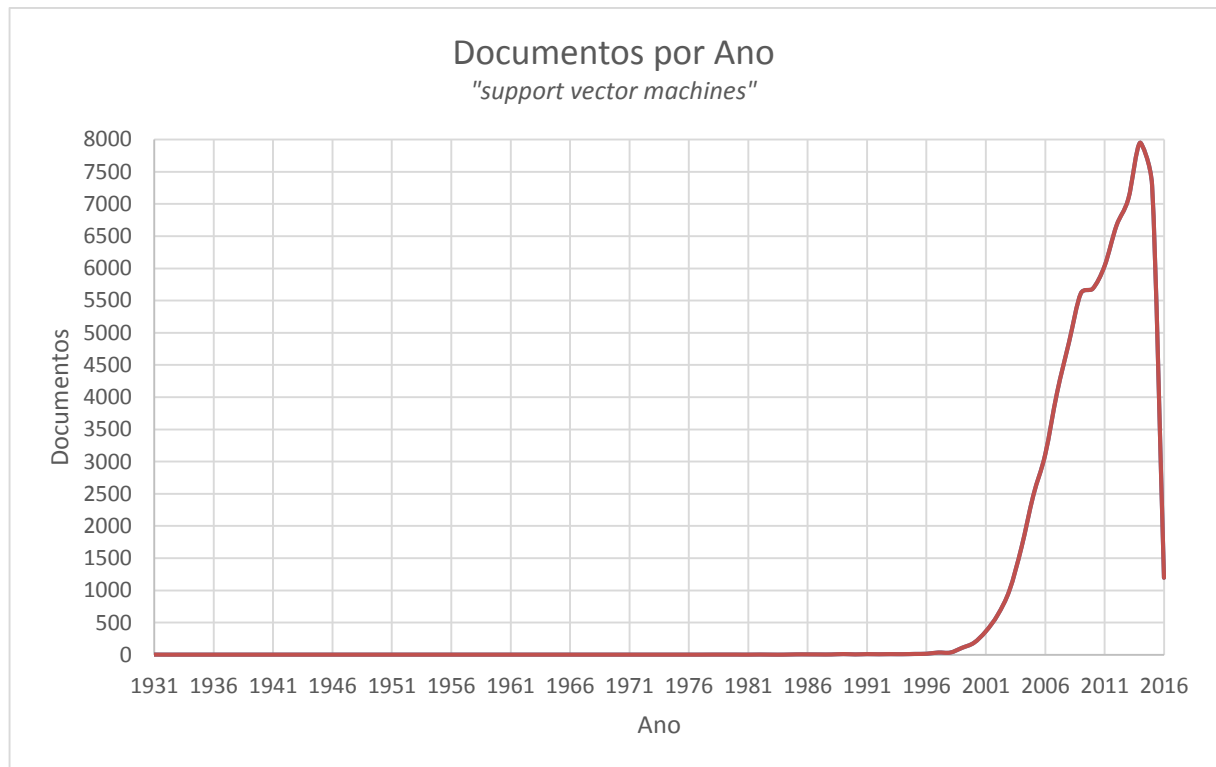
As máquinas de vetores suporte (*SVM – Support Vector Machines*) são algoritmos de aprendizagem de máquina baseados na Teoria do Aprendizado Estatístico (CORTES; VAPNIK, 1995) e têm sido introduzidas recentemente com sucesso na Estatística Multivariada (PENG; LI, 2014). Em aplicações diversas (CUI; YAN, 2009; NADADOOR et al., 2012; SHENG et al., 2013) na resolução de problemas de classificação (LUTS et al., 2010) e de regressão, na presença de não-linearidades (BALABIN; SMIRNOV, 2011), têm apresentado desempenho superior aos algoritmos de referência tais como Análise por Componentes Principais (*PCA – Principal Component Analysis*), Modelagem Independente e Flexível por Analogia de Classe (*SIMCA – Soft Independent Modeling of Class Analogy*) (BALABIN; SAFIEVA; LOMAKINA, 2010; ALVES, HENRIQUES, POPPI, 2014), Mínimo Quadrados Parciais para Análise Discriminate (*PLS-DA – Partial Least Squares Discriminant Analysis*) (YIP et al., 2014) e Regressão por Mínimos Quadrados Parciais (*PLSR – Partial Least Squares Regression*) (LI; LIANG; XU, 2009).

Considerando-se as questões levantadas na literatura sobre as máquinas de vetores suporte (SVM) em problemas de regressão na área química, esta pesquisa apresenta várias contribuições tecnológicas e acadêmicas proporcionando significativos avanços científicos, uma vez que modificaria a maneira de se pensar no que se diz respeito a métodos experimentais para monitoramento de processos industriais em geral.

1. Agilidade no processo de tomada de decisão;
2. Utiliza técnicas não destrutivas;
3. Custo de implementação relativamente baixo em relação ao retorno financeiro para o negócio;
4. Necessita de pequenas quantidades de amostra;
5. Não gera resíduos químicos;
6. Elimina reprocessamento de produtos, o que reduz o impacto sobre o meio ambiente;
7. Reduz a quantidade de ensaios físico-químicos no laboratório de qualidade.
8. São escassos os trabalhos na literatura sobre aplicações de SVM para a predição de parâmetros de qualidade de produtos químicos acabados.
9. A literatura apresentou poucos estudos na otimização de parâmetros do PLS via projeto experimental de Taguchi nesta área de estudo, bem como Taguchi como método de seleção de variáveis.

Nos últimos anos, o número de trabalhos envolvendo *SVM* vem aumentando como pode ser visto na Figura 1 representando um total de 66095 resultados, o que prova a exploração desta técnica pela comunidade científica.

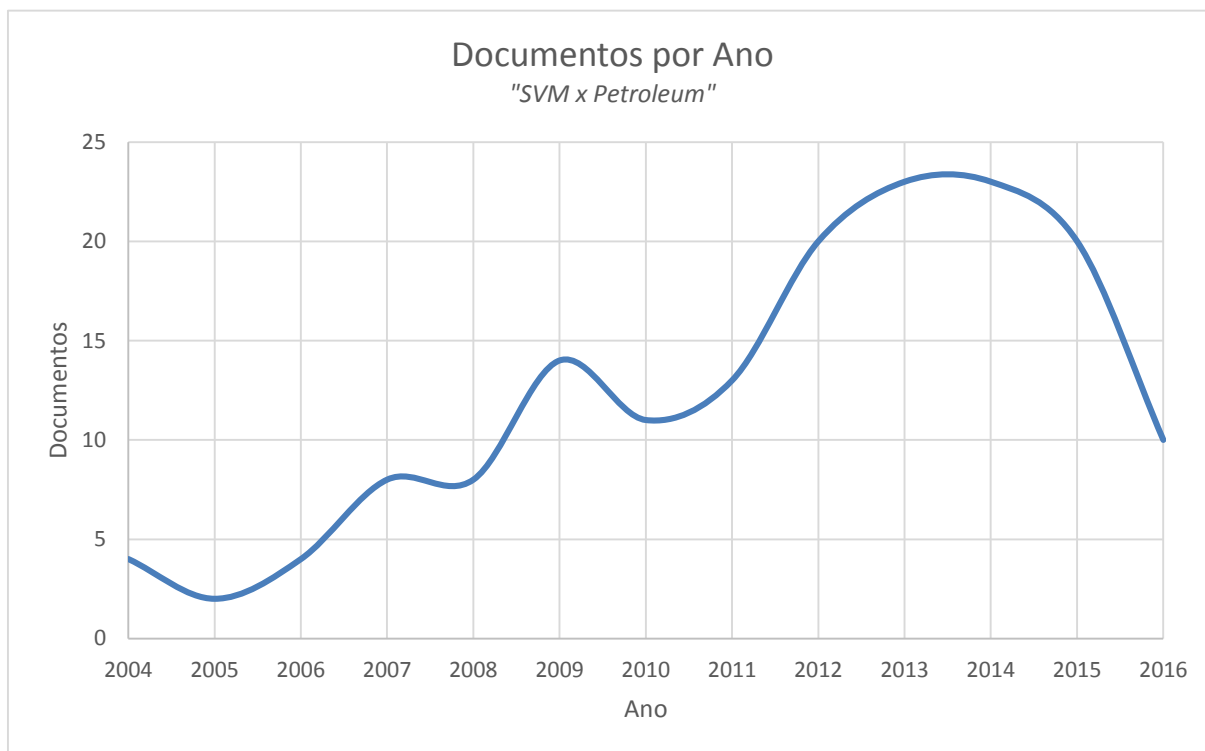
Figura 1 – Publicações com relação à palavra-chave *SVM*.



Fonte: Scopus

Relacionando as palavras-chaves *SVM* x *matéria-prima* foi possível verificar a ocorrência de 160 documentos de 2004 a 2017. Dentre eles, 35,0% estão relacionadas à área das Engenharias, 25,6% na área de Energia e 22,5% na área de Ciência da Computação conforme Figura 2.

Figura 2 – Publicações com relação às palavras-chave *SVM x produto*



Fonte: Scopus

Apesar da crescente utilização desse campo de pesquisa, não foi possível ainda encontrar na literatura trabalhos relacionados à utilização de *LS-SVM* para previsão de propriedades físico-químicas da matéria-prima conforme Tabela 1.

Tabela 1 - Ocorrências das palavras-chave nas bases de dados *Scopus* e *Web of Science*.

Base de Dados	Palavras-Chave	Ocorrência
<i>Scopus</i>	<i>SVM x fuel</i>	160
	<i>LS-SVM x fuel</i>	11
	<i>LS-SVM x fuel</i>	0
	<i>LS-SVM Taguchi x fuel</i>	0
	<i>LS-SVM x Taguchi x fuel</i>	0
<i>Web of Science</i>	<i>SVM x fuel</i>	48
	<i>LS-SVM x fuel</i>	4
	<i>LS-SVM x fuel</i>	0
	<i>LS-SVM Taguchi x fuel</i>	0

Fonte: Scopus e Web of Science

1.4 Delimitações do trabalho

As delimitações do trabalho desenvolvido nesta tese são:

- Os resultados dos modelos empíricos obtidos são válidos apenas para o produto orgânico estudado, embora, a abordagem utilizada possa ser estendida a qualquer produto químico sensível à radiação infravermelha.
- A presente tese visa comparar as máquinas de vetores suporte com o método tradicional PLS, amplamente utilizado na literatura, e não com outras estratégias de tratamento de dados não lineares.
- Não é escopo do trabalho otimizar os processos industrial em si e, sim disponibilizar modelos capazes de prever as propriedades físico-químicas do produto final que futuramente poderão pautar a otimização do processo produtivo.

1.5 Método e classificação da Pesquisa

O método de pesquisa utilizado em todas as etapas do trabalho foi a Pesquisa-Ação. De acordo com Miguel et al. (2010) este trabalho pode ser classificado como de natureza aplicada, com objetivos axiomáticos normativos e de abordagem quantitativa.

As fases da pesquisa conforme Figura 3, foram conduzidas segundo a estrutura da metodologia Seis Sigma, definida por Linderman et al. (2003) como um método organizado e sistemático para a melhoria de processos e desenvolvimento de novos produtos e serviços, que se baseia em métodos estatísticos e científicos para fazer reduções drásticas nas taxas de defeitos para o cliente. Este procedimento geralmente conhecido como DMAIC (*Define, Measure, Analyse, Improve, Control*) – Definir, Medir, Analisar, Melhorar e Controlar (PUGNA; NEGREIA; MICLEA, 2016)

Passa-se a descrever, sucintamente, cada uma das etapas da pesquisa.

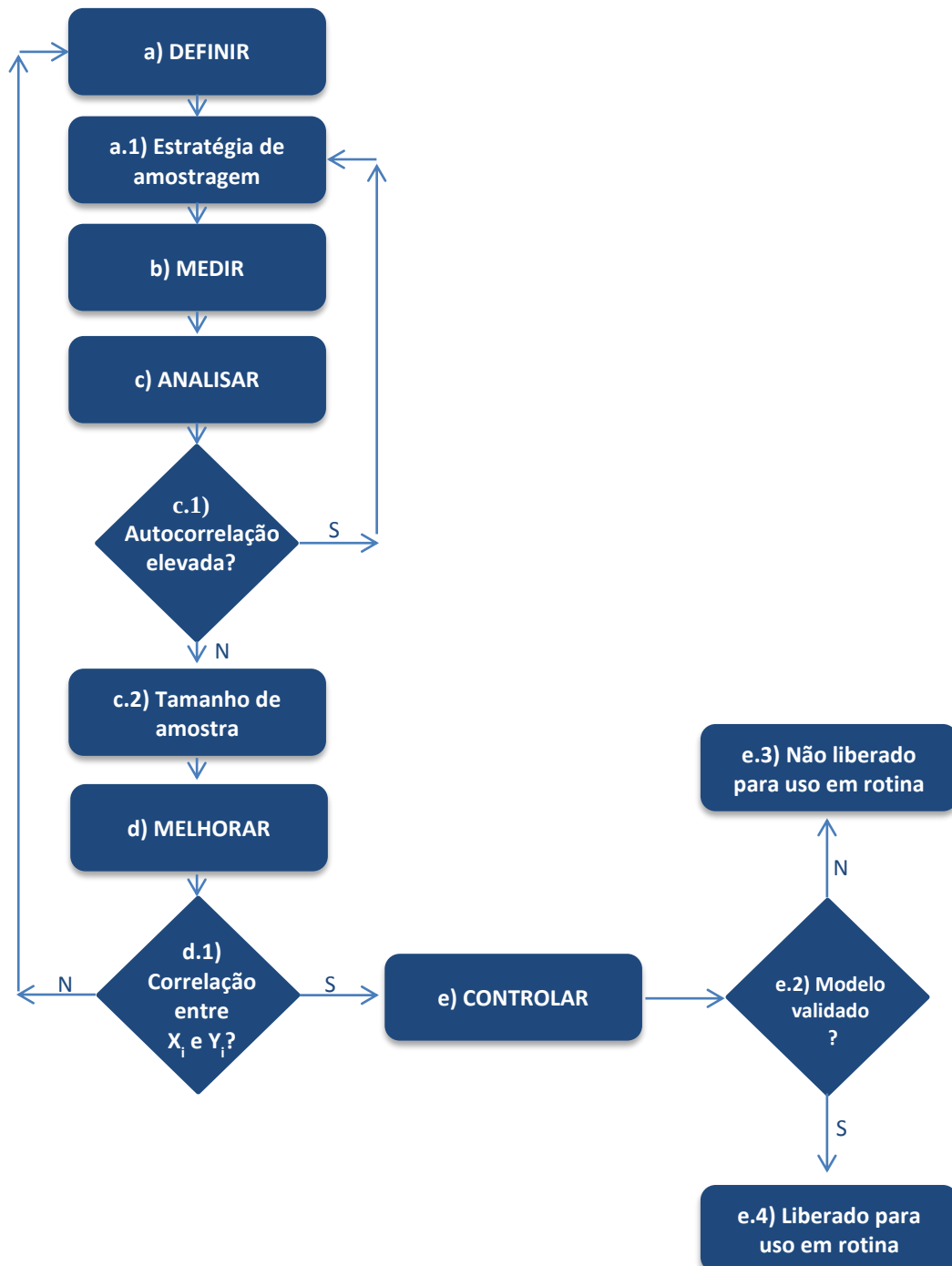
(a) Fase Definir:

Em uma indústria química foi identificado o problema a ser estudado que consiste no desenvolvimento de modelos multivariados de predição de propriedades físico-químicas com vistas ao monitoramento da produção.

As variáveis envolvidas neste sistema tanto para SVM quanto para PLS são:

- Variáveis de entrada (X 's) ou preditoras: absorvância de radiação infravermelha em função de números de onda (cm^{-1}) do produto químico em estudo. Esta técnica analítica será detalhada no Capítulo 2.

Figura 3 – Etapas da Pesquisa



Fonte: Produção do próprio autor

- Variáveis de saída (Y's): são as propriedades físico-químicas da amostra - propriedades P1, P2, P3, P4 e P5 (°C); P6(%); P7 (°C); P8 (%) e P9 a -20°C (mm²/s).

Com respeito à estratégia de amostragem, passa-se a descrever o que foi realizado. Processos químicos industriais, naturalmente produzem dados autocorrelacionados à partir de observações sucessivas em intervalos curtos de tempo de uma mesma característica de qualidade. (COSTA; EPPRECHT; CARPINETTI, 2010), o que levou a necessidade de se planejar coletas em momentos em que a indústria produzisse amostras independentes umas das outras.

Esta variabilidade é possível após troca da matéria-prima da indústria química em estudo, pois misturas de produtos diferentes possuem composições químicas diferentes e, portanto, amostras coletadas após troca de carga são estatisticamente distintas umas das outras (CAMOLESI, 2009).

De acordo com Costa, Epprecht e Carpinetti (2010) atualmente, processos industriais contínuos, cada vez mais automatizados, como os químicos, costumam produzir dados autocorrelacionados, isto é, com um grau de correlação entre observações sucessivas X_i e X_{i-k} de uma mesma característica de qualidade, onde k indica a defasagem das observações. O coeficiente de autocorrelação amostral pode ser calculado pela Equação 1:

$$r_k = \frac{\sum_{i=k+1}^n (x_i - \bar{X})(x_{i-k} - \bar{X})}{\sum_{i=1}^n (x_i - \bar{X})^2} \quad (1)$$

(b) Fase Medir:

Nesta etapa, as amostras químicas utilizadas foram submetidas a testes em laboratório. As propriedades físico-químicas foram medidas segundo métodos de referência da ASTM (ASTM – *American Standard for Testing and Materials*). As mesmas amostras foram submetidas à radiação infravermelha.

(c) Fase Analisar:

Nesta fase é observada a presença de variabilidade dos dados em X e Y. Caso ocorram falhas nas amostragens a probabilidade de formar um banco de dados com amostras estatisticamente semelhantes é alta, principalmente se as amostras forem coletadas em intervalos curtos de tempo dentro de uma mesma campanha de produção. A variabilidade em X pode ser verificada por meio de Análise por Componentes Principais (PCA – *Principal Component Analysis*) e em Y por meio do perfil de histogramas conforme Capítulo 3.

É fundamental formar um banco de dados com alta variabilidade (c.1) para que o conjuntos de modelagem possam treinar o maior número possível de perfis de produção contribuindo desta forma com a capacidade de generalização de amostras futuras. Em caso de sucesso seguir para a fase (d), caso contrário, voltar à etapa (a.1) e revisar a sistemática de amostragem.

Esta etapa também foi dedicada à definição do tamanho de amostra (c.2) para formação dos conjuntos de treinamento e validação externa. O critério proposto foi o algoritmo *y-rank* que classifica os dados em ordem crescente em relação ao vetor de respostas Y. A partir de uma razão de segmentação, o algoritmo distribui os dados em dois subconjuntos. Neste trabalho, sem descontar possíveis, *outliers*, foi utilizada a razão de segmentação 3:1, ou seja, a cada três amostras para o conjunto de treinamento, uma para o conjunto de validação externa (FACCHIN, 2005; (RANZAN, 2014).

(d) Fase Melhorar

Nesta etapa, foram desenvolvidos os modelos de predição SVM e PLS utilizando a biblioteca gráfica *The Unscrambler*.

Foi também proposto um algoritmo de otimização dos modelos PLS, o *PLS-Taguchi-Desirability*, que buscou auxiliar a configuração desta modelagem mediante um problema de otimização de múltiplas respostas com objetivo de contornar o sobreajuste do dados. Neste contexto, foi proposta a utilização conjunta do arranjo ortogonal L_9 de Taguchi e *Desirability* para otimização dos parâmetros de modelagem do PLS.

(d.1) Há correlação entre X_i e Y_i ?

Caso não haja correlação entre as variáveis preditoras X_i e as variáveis respostas Y_i , é necessário voltar à etapa 1 e verificar as falhas na identificação do problema e nos métodos propostos e, inclusive se o infravermelho tem sensibilidade à propriedade a ser modelada. Caso contrário, seguir para a fase (e).

(e) Fase Controlar:

Modelo Validado? O desempenho dos modelos foi avaliado por meio de validação interna e externa, seguida do teste F de *Fisher* para comparação da equivalência estatística dos erros de predição para os conjuntos de treinamento e validação externa.

Se o desempenho não for considerado satisfatório, seguir para a etapa (e.4), sendo assim, pode ser utilizado nos processos, caso contrário, seguir para etapa (e.3) representando a não liberação dos modelos para uso em rotina. Ainda que um modelo multivariado apresente correlação entre as variáveis preditoras X_i e variáveis respostas Y_i pode ser considerado não validado, se o erro padrão de validação externo for estatisticamente não equivalente ao erro padrão de validação interna, indicando baixa robustez na generalização de amostras futuras.

1.6 Estrutura do Trabalho

Esta tese será dividida em mais três capítulos, além deste introdutório.

O Capítulo 2 traz aplicações da Análise Multivariada, em seguida, tratará da ciência e tecnologia envolvidas na mineração dos dados. Posteriormente, será explorado o estudo dos métodos multivariados utilizados.

O Capítulo 3 apresentará os resultados obtidos. Os tratamentos estatísticos foram realizados pelos softwares *The Unscrambler*® versão X.10.4.1, *Minitab*® versão 14, *STATISTICA*® versão 7 e Excel® versão 2013.

No Capítulo 4, estão as considerações finais da tese e recomendações futuras seguidas dos Anexos e referências bibliográficas utilizadas neste texto.

4 CONCLUSÕES

4.1 Verificação dos objetivos e respostas às questões de pesquisa

Foi possível modelar matematicamente as propriedades P1, P2, P3, P4, P5, P6, P7, P8 e P9 pela técnica multivariada SVM a partir de dados de absorção da radiação infravermelha.

O fenômeno da multicolinearidade espectral foi comprovado pela amplitude dos coeficientes de correlação a cada dois números de onda dos espectros, o que confirma a necessidade de trabalhar a dimensionalidade dos dados para desenvolver equações de regressão já que a regressão tradicional MLR não seria possível.

Através da demonstração da amplitude dos coeficientes de autocorrelação amostral, foi observada a eficiência da estratégia de amostragem após troca de matéria-prima de uma indústria química, contornando essa questão intrínseca do processo químico industrial.

O algoritmo proposto PLS–Taguchi seguido de otimização via *Desirability* permitiu identificar a melhor configuração possível do PLS.

Foi observada a robustez nos modelos obtidos via máquinas de vetores suporte, que trabalham com otimização convexa, devido à equivalência estatística dos indicadores RMSECV e RMSEP e, compatibilidade à reprodutibilidade dos métodos ASTM de referência.

4.2 Sugestões para continuidade do trabalho

Esta pesquisa pode ter continuidade em projetos outros projetos de doutorado ou pós-doutorado. Recomenda-se:

- 1) Construir cartas de controle estatístico de processos para valores individuais e média móvel, I – MR (*Individual, Moving Range*) para validação contínua dos modelos de predição, isto é, estabelecer limites de controle para os resíduos entre os valores preditos pelos modelos e valores dos métodos de referência, com isso, sinais estatísticos irão indicar os momentos para se fazer atualizações nos modelos e, consequentemente, garantir sua robustez ao longo do tempo;
- 2) Testar métodos de otimização alternativo ao *Desirability*;
- 3) Comparar as máquinas de vetores suporte com as redes neurais artificiais e redes otimizadas por Planejamento de Experimentos com foco na convexidade da função objetivo;
- 4) Testar outros algoritmos na designação dos conjuntos de treinamento e validação tais como o Algoritmo de *Kennard-Stone* que se baseia em distância estatística levando-se em consideração a matriz X de dados espectrais.
- 5) Testar outros algoritmos para seleção de variáveis do espectro de infravermelho;

5 REFERÊNCIAS BIBLIOGRÁFICAS

- ABE, S. In: **Support vector machine for pattern classification**. London. Springer, 343p, 2005.
- ALVES, J. C.; HENRIQUES, C. B.; POPPI, R. J. Determination of diesel quality parameters using support vector regression and near infrared spectroscopy for in-line blending optimizer system. **Fuel**, v.97, p. 710-717, 2012.
- ALVES, J. C.; HENRIQUES, C. B.; POPPI, R. J. Classification of diesel pool refinery streams through near infrared spectroscopy and support vector machines using C-SVC and v-SVC. **Analytica Chimica Acta**, v.117, p. 389-396, 2014.
- ALVES, J. C.; POPPI, R. J. Biodiesel content determination in diesel fuel blends using near infrared (NIR) spectroscopy and support vector machines (SVM). **Talanta**, v.104, p. 155-161, 2013.
- AMARAL, F. C. N. Data mining – técnicas e aplicações para o marketing direto. In: **Data Mining**, São Paulo, Berkeley, p1-110, 2001.
- ANDERSON, T. W. **An Introduction to Multivariate Statistical**, 3 ed. Hoboken Wiley Series in Probability and Statistics, 2003. 742p.
- ANDRADE, J. M.; GARCIA, M. V.; MAHIA, P. L.; PRADA, D. A review of the main factors influencing the FT-IR-PLS abilities exemplified with petrochemical qualimetric applications. **Talanta**, v.44, p. 2167-2184, 1997.
- BAFFI, G.; MARTIN, E. B.; MORRIS, A. J. Non-linear projection to latent structures revisited (the neural network PLS algorithm). **Computers and Chemical Engineering**, v. 23, p. 1283-1307, 1999.
- BALABIN, R. M.; SAFIEVA, R. Z.; LOMAKINA, E. Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction. **Chemometrics Intelligent Laboratory Systems**, v.88, p. 183-188, 2007.
- BALABIN, R. M.; SAFIEVA, R. Z. Gasoline classification by source and type based on near infrared (NIR) spectroscopy data. **Fuel**, v.87, p. 1096-1101, 2008.
- BALABIN, R. M.; SAFIEVA, R. Z.; LOMAKINA, E. Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques. **Analytica Chimica Acta**, v.671, p. 27-35, 2010.
- BALABIN, R.; LOMAKINA, E. I.; SAFIEVA, R. Z. Neural network (ANN) approach to biodiesel density, kinematic viscosity, methanol and water contents using near infrared (NIR) spectroscopy. **Fuel**, v.90, p. 2007-2015, 2011.
- BALABIN, R. M.; SMIRNOV, S. V. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. **Analytica Chimica Acta**, v.692, p. 63-72, 2011.

BARROS, A. S.; RUTLEDGE, D. N. Principal components transform-partial least squares: a novel method to accelerate cross-validation in PLS regression. **Chemometrics and Intelligent Laboratory Systems**, v. 73, p. 245-255, 2004.

BASSBASI, M.; HAFID, A.; PLATIKANOV, S.; TAULER, R.; OUSSAMA, A. Study of motor oil adulteration by infrared spectroscopy and chemometrics methods. **Fuel**, v.104, p. 798-804, 2013.

BASTIEN, P.; VINZI, V. E.; TENENHAUS, M. PLS generalised linear regression. **Computational Statistics & Data Analysis**, v. 48, p. 17-46, 2005.

BEHIN, J.; FARHADIAN, N. Multi-objetive optimization of oxidative desulfurization in a sono-photochemical airlift reactor. **Ultrasonics Sonochemistry**, v. 38, p. 50-61, 2017.

BERGER, P. D.; MAURER, R. Introduction to Taguchi Methods. In: **Experimental Design with Applications in Management, Engineering and the Sciences**. United States. Duxbury Thomson Learning, 479 p., 2002.

BI, Y.; YUAN, K.; XIAO, W.; WU, J.; SHI, C.; XIA, J.; CHU, G.; ZHANG, G.; ZHOU, G. A local pre-processing method for near-infrared spectra, combined with spectral segmentation and standard normal variate transformation. **Analytica Chimica Acta**, v. 909, p. 30-40, 2016.

BRAGA, J. W.; JUNIOR, A. A. S.; MARTINS, I. S. Determination of viscosity index in lubricant oils by infrared spectroscopy and PLSR. **Fuel**, v. 120, p. 171-178, 2014.

CAMOLESI, V. J. **Caracterização do querosene através da espectroscopia de infravermelho próximo**. 2009. 50 f. Dissertação (Mestrado em Engenharia) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2009.

CHAI, J.; LIU, J. N. K.; NGAI, E. W. T. Application of decision-making techniques in supplier selection: A systematic review of literature. **Expert Systems with Applications**, v.40, p. 3872-3885, 2013.

CHEN, C.; HUANG, C. Improve electromagnetic interference of electronic products with Taguchi parametric design. **Measurement**, v.102, p. 200-207, 2017.

CHOU, J. S.; CHENG, M. Y.; WU, Y. W.; PHAM, A D. Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classifications. **Expert Systems with Applications**, v.41, p. 3955-3964, 2014.

COELHO, M. B.; COELHO, J.; ITURRIAGA, H.; MASPOCH, S.; PAGÈS, J. Calibration in non-linear near infrared reflectance spectroscopy: a comparison of several methods. **Analytica Chimica Acta**, v.384, p. 207-214, 1999.

CORTES, C.; VAPNIK, V. Support-Vector Networks. **Machine Learning**, v.20, p.273-297, 1995.

COSTA, A. F. B.; EPPRECHT, E. K.; CARPINETTI, L.C. R. Introdução e conceitos fundamentais. In: **Controle Estatístico de Qualidade**. São Paulo. Atlas, p15-20, 2010.

CUI, W.; YAH, X. Adaptive wighted least square support vector machine regression integrate with outlier detection and its application in QSAR. **Chemometrics and Intelligent Laboratory Systems**, v.98, p. 130-135, 2009.

DEVOS, O.; DOWNEY, G.; DUPONCHEL, L. Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils. **Food Chemistry**, v.148, p. 124-130, 2014.

DU, S.; LV, J.; XI, L. On-line classifying process mean shifts in multivariate control charts based on multiclass support vector machines. **International Journal of Production Research**, v.50, p. 6288-6310, 2012.

DUARTE, L. M.; FILGUEIRAS, P. R.; SILVA, S. R. C.; DIAS, J. C. M.; OLIVEIRA, L. M. S. L.; CASTRO, E. V. R.; OLIVEIRA, M. A. L. Determination of some physicochemical properties in Brazilian crude oil by ^1H NMR spectroscopy associated to chemometric approach. **Fuel**, v. 181, p. 660-669, 2016.

ERGON, R. Informative PLS score-loading plots for process understanding and monitoring. **Journal of Process Control**, v. 14, p. 889-897, 2004.

FACCHIN, S. **Técnicas de Análise Multivariável Aplicadas ao Desenvolvimento de Analisadores Virtuais**. 2005. 72 f. Dissertação (Mestrado em Engenharia) – Departamento de Engenharia Química, Universidade Federal do Rio Grande do Sul, UFRS, 2005.

FALLA, F. S. F.; LARINI, C; LE ROUX, G. A. C.; QUINA, F. H.; MORO, L. F. L.; NASCIMENTO, C. A. O. Characterization of crude petroleum by NIR. **Petroleum Science & Engineering**, v.51, p. 127-137, 2006.

FAYAZI, A.; ARABLOO, M.; MOHAMMADI, A. H. Efficient estimation of natural gas compressibility factor using a rigorous method. **Journal of Natural Gas Science and Engineering**, v.16, p. 8-17, 2014.

FERREIRA, M. M. C.; MONTANARI, C. A.; GAUDIO, A. C. Seleção de Variáveis em QSAR. **Química Nova**, v. 25, p. 439-448, 2002.

FILGUEIRAS, P. R.; ALVES, J. C.; POPPI, R. J. Quantification of animal fat biodiesel in soybean biodiesel and B20 diesel blends using infrared spectroscopy and synergy interval support vector regression. **Talanta**, v.119, p. 582-589, 2014.

FILGUEIRAS, P. R.; SAD, C. M. S.; LOUREIRO, A.; SANTOS, M. F. P.; CASTRO, E. V. R. Determination of API gravity, kinematic viscosity and water content in petroleum by ATR-FTIR spectroscopy and multivariate calibration. **Talanta**, v.119, p. 582-589, 2014.

FILGUEIRAS, P. R.; TERRA, L. A.; CASTRO, E. V. R.; OLIVEIRA, L. M. S. L.; DIAS J. C. M.; POPPI, R. J. Prediction on the distillation temperatures of crude oils using ^1H NMR and

support vector regression with estimated confidence intervals. **Talanta**, v. 142, p. 197-205, 2015.

FORATO, L. A.; FILHO, R. B.; COLNAGO, L. A. Estudos de métodos de aumento de resolução de espectros de FTIR para análises de estruturas secundárias de proteínas. **Química Nova**, v. 20, p. 146-150, 1997.

FREITAS, L. V. **Análise multivariada no monitoramento do processo de produção de óleo diesel numa refinaria de petróleo**. 2012. 51 f. Dissertação (Mestrado em Engenharia Mecânica) – Departamento de Produção, Universidade Estadual Paulista, UNESP, 2012.

FREITAS, L. V.; FREITAS, A. P. B. R.; MARINS, F. A.; VERASZTO, E. V.; SILVA, M. B. Partial least squares for prediction the physicochemical properties in petroleum products. **European International Journal of Science and Technology**, v.2, p.87-92, 2013.

GANI, W.; TALEB, H.; LIMAM, M. Support vector regression based residual control charts. **Journal of Applied Statistics**, v.2, p. 309-324, 2010.

GELADI, P.; KOWALSKI, B. Partial Least Square Regression: A Tutorial. **Analytica Chimica Acta**, v. 185, p. 1-17, 1986.

GUPTA, A. K. Predictive modelling of turning operations using response surface methodology, artificial neural networks and support vector regression. **International Journal of Production Research**, v.48, p. 763-778, 2010.

LI, H.; LIANG, Y.; XU, Q. Support vector machines and its applications in chemistry. **Chemometrics and Intelligent Laboratory Systems**, v.95, p. 188-196, 2009.

LI, D. C.; LIU, C. W.; HU, S. A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. **Artificial Intelligence in Medicine**, v.52, p. 45-52, 2011.

LI, X.; CHEN, F.; LI, S.; JIA, J.; GU, H.; YANG, L. An efficient homogenate-microwave-assisted extraction of flavonols and anthocyanins from blackcurrant marc: Optimization using combination of Plackett-Burman design and Box-Behnken design. **Industrial Crops and Products**, v.94, p. 834-847, 2016.

GALVÃO, R. K. H.; ARAÚJO, M. C. U.; MARTINS, M. N.; JOSÉ, G. E.; PONTES, M. J. C.; SILVA, E. C.; SALDANHA, T. C. B. An application of subagging for the improvement of prediction accuracy of multivariate calibration models. **Chemometrics and Intelligent Laboratory Systems**, v. 81, p. 60-67, 2006.

GALVÃO, R. K. H.; ARAÚJO M. C.; FRAGOSO, W. D.; SILVA, E. C.; JOSÉ, G. E.; SOARES, S. F. C.; PAIVA, H. M. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. **Chemometrics and Intelligent Laboratory Systems**, v. 92, p. 83-91, 2008.

GONZAGA, F. B.; PASQUINI, C. A low cost short wave near infrared spectrophotometer: Application for determination of quality parameters of diesel fuel. **Analytica Chimica Acta**, v.670, p. 92-97, 2010.

HARIHARAN, M.; POLAT, K.; SINDHU, R. A new hybrid intelligent system for accurate detection of Parkinson's disease. **Computers Methods and Programs in Biomedicine**, v.113, p.904-913, 2014.

HONG, E. C.; DONG, Y.; ZHENG, F.; LAI, C.Y. Forecasting urban traffic flow by SVR continuous ACO. **Applied Mathematical Modelling**, v.35, p. 1282-1291, 2011.

HONORATO, F. A.; NETO, B. B.; PIMENTEL, M. F.; STRAGEVITCH, L.; GALVÃO, R. K. H. Using principal component analysis to find the best calibration settings for simultaneous spectroscopic determination of several gasoline properties. **Fuel**, v. 87, p. 3706-3709, 2008.

HARDING, J.A.; SHAHBAZ, M.; SRINIVAS, S; KUSIAK, A. Data mining in manufacturing: a review. **Journal of Manufacturing Science and Engineering**, v. 128, p. 969 – 976, 2006

HENRY, R. C.; PARK, E. S.; SPIEGELMAN, C. H. Comparing a new algorithm with the classic methods for estimating the number of factors. **Chemometrics and Intelligent Laboratory Systems**, v. 48, p. 91-97, 1999.

HERNÁNDEZ, N.; TALAVERA, I.; BISCAY, R. J.; PORRO, D.; FERREIRA, M. M. C. Support vector regression for functional data in multivariate calibration problems. **Analytica Chimica Acta**, v.642, p. 110-115, 2009.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**, 5 ed. Madinson: Prentice-Hall, 2002. 797p.

JOLLIFFE, I. T. **Principal Component Analysis**, 2 ed. New York. Springer Series in Statistics, 2002. 350p.

KANDETHODY, M.; TSOKS, C. P. Chapter 9 – Design of Experiments. **Mathematical Statistics with Applications in R (Second Edition)**, p. 459-494, 2015.

KARDAMAKIS, A. A.; PASADAKIS, N. Autoregressive modeling of near-IR spectra and MLR to predict RON values of gasolines. **Fuel**, v. 89, p. 158-161, 2010.

KHANMOHAMMADI, M.; GARMARUDI, A. B.; GUARDIA, M. Characterization of petroleum-based products by infrared spectroscopy and chemometrics. **Trends in Analytical Chemistry**, v. 35, p. 135-149, 2012.

KHASHAN, M. A.; NAGGAR, A. M. E.; SHADDAD, E. A new method of determining the optical constants of a thin film from its reflectance and transmittance interferograms in a wide spectral range: 0.2–3 μm . **Optics Communications**, v. 178, p. 123-132, 2000.

KILLNER, M. H. M.; ROHWEDDER, J. J. R.; PASQUINI, C. A PLS regression model using NIR spectroscopy for on-line monitoring of the biodiesel production reaction. **Fuel**, v.90, p. 3268-3273, 2011.

KOWALSKA, A.; MACIEJEWSKI, K. Multivariate optimization of recycled road base cold mixtures with foamed bitumen. **Procedia Engineering**, v. 108, p. 436-444, 2015.

KRANJC, J.; ORAC, R.; PODPECAN, V.; LAVRAC, N.; SIKONJA, M. ClowdFlows: Online workflows for distributed big data miming. **Future Generation Computer Systems**, v.68, p. 38-58, 2017.

KIM, K.; CHO, I.; PARK, J. Use of real-time NIR (near infrared) spectroscopy for the on-line optimization of a crude distillation unit. In: NPRA, **Computer Conference**. Chicago, 2000.

KIRSANOV, D.; MEDNOVA, O.; VIETORIS V.; KILMARTIN P. A.; LEGIN, A. Towards reliable estimation of an “electronic tongue” predictive ability from PLS regression models in wine analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 90, p. 109-116, 2012.

KUANG, F.; XU, W.; ZHANG, S. A novel hybrid KPCA and SVM with GA model for intrusion 97fluente97. **Applied Soft Computing**, v.18, p. 178-185, 2014.

LAMA, N.; BORACCHI, P.; BIGANZOLI, E. Partial logistic relevance vector machines in survival analysis. **Journal of Applied Statistics**, v.11, p.2445-2458, 2011.

LAU, C. K.; LAI, K. K.; LEE, Y. P. DU, J. Fire risk assessment with scoring system, using the support vector machine approach. **Fire Safety Journal**, v. 78, p. 188-195, 2015.

LAURENTINO, J. C. **Máquina de vetores suporte aplicada a dados de espectroscopia NIR de combustíveis e lubrificação para o desenvolvimento de modelos de regressão e classificação**. 2012. 136 f. Tese (Doutorado em Ciências) – Instituto de Química, Universidade Estadual de Campinas, UNICAMP, 2012.

LINDERMAN, K.; SCHROEDER, R.; ZAHEER, S.; CHOO, A. Six Sigma: a goal-theoretic perspective. **Journal of Operations Management**, v. 21, p. 193-203, 2003.

LIU, X.; ZHANG, Z.; LIANG, Y.; SOUSA, P.; YUN, Y.; YU, L. Baseline correction of high resolution spectral profile data based on exponential smoothing. **Chemometrics and Intelligent Laboratory Systems**, v. 139, p. 97-108, 2014.

LIRA, L. F. B.; VASCONCELOS, F. V. C.; PEREIRA, C. F.; PAIM, A. P. S.; STRAGEVITCH, L.; PIMENTEL M. F. Prediction of properties of diesel/biodiesel blends by infrared spectroscopy and multivariate calibration. **Fuel**, v. 89, p. 405-409, 2010.

LU, C. J.; WANG, Y. W. Combining independent component analysis and growing hierarchical self-organizing maps with support vector regression in product demand forecasting. **International Journal of Production Economics**, v.128, p.603-613, 2010.

LUTS, J.; OJEDA, F.; PLAS, R. V.; DE MOOR, B.; HUFFEL, S. V.; SUYKENS, J. A. K. A tutorial on support vector machine-based methods for classification problems in chemometrics. **Analytica Chimica Acta**, v.665, p. 129-145, 2010.

MABOOD, F.; GILANI, S.; ALBROUMI, M.; ALAMERI, S.; NABBANI, M.; JABEEN, F.; HUSSAIN, J.; HARRASI, A.; BOQUÉ, R.; FAROOQ, S.; HAMAED, A.; NAUREEN, A.; KHAN, A.; HUSSAIN, Z. Baseline correction of high resolution spectral profile data based on exponential smoothing. **Chemometrics and Intelligent Laboratory Systems**, v. 139, p. 97-108, 2014.

MACHADO, M. A. G.; COSTA, A. F. B. The use of principal components and univariate charts to control multivariate processes. **Pesquisa Operacional**, v. 28, p. 173-196, 2008.

MALDONADO, S.; PÉREZ J.; BRAVO, C. Cost-based feature selection for support vector machines: an application in credit scoring. **European Journal of Operational Research**, v.261, p. 656-665, 2017.

MANDAL, S. K.; CHAN, F. T. S.; TIWARI, M. K. Leak detection of pipeline: An integrated approach of rough set theory and artificial bee colony trained SVM. **Expert Systems with Applications**, v.39, p. 3071-3080, 2012.

MAO, W.; YAH, G.; DONG, L.; H, D. Model selection for least squares support vector regressions based on small-strategy. **Expert Systems with Applications**, v.38, p. 3227-3237, 2011.

MARJANOVIC, M.; KOVACEVIC, M.; BAJAT, B.; VOZENÍLEK, V. Landslide susceptibility assessment using SVM machine learning algorithm. **Engineering Geology**, v.123, p. 225-234, 2011.

MARTINS, J. P. A.; TEÓFILO, R. F.; FERREIRA, M. M. C. Computational performance and cross-validation error precision of five PLS algorithms using designed and real data sets. **Journal of Chemometrics**, v. 24, p. 320-332, 2010.

MENG, X.; JIA, M.; WANG, T. Neural network prediction of biodiesel kinematic viscosity at 313 K. **Fuel**, v.121, p. 133-140, 2014.

MIGUEL, A. C. M; FLEURY, A; MELLO, C. H. P; NAKANO, D. N; TURRIONI, J. B; HO, L. L; MORABITO, R; MARTINS, R. A; PUREZA, V. **Metodologia de pesquisa em engenharia de produção e gestão de operações**. 1 ed. Rio de Janeiro. Editora Campus, 2010, 248p.

MINGOTI, S. **Análise de dados através de métodos de estatística multivariada: Uma abordagem aplicada**. 1. ed. Belo Horizonte: Editora UFMG, 2005. 297p.

MONEDERO, I.; BISCARRI, F.; LEÓN, C.; GUERRERO, J. I.; GONZÁLEZ, R.; LOMBARD, L. P. Decision system based on neural networks to optimize the energy efficiency of petrochemical plant. **Expert Systems with Applications**, v.39, p. 9860-9867, 2012.

MOHANTY, S.; MISHRA, A.; NANDA, B.; ROUTARA Multi-objective parametric optimization of nano poder mixed electrical discharge machining of AlSiCp using reponse surfasse methodology and particle swarm optimization. **Alexandria Engineering Journal**, Article in press, 2017.

MONTGOMERY, D. C. In: **Design and Analysis of Experiments**. John Wiley & Sons, 5 ed., 699p.

MONTGOMERY, D. C.; RUNGER, G. C. Estatística Aplicada e Probabilidade para Engenheiros, LTC – Livros Técnicos e Científicos Editora AS, 2 ed., 570p.

MORADI, S.; ALAM, M. Multi-criteria optimization of lateral load-drift response of posttensioned steel beam-coloumn connections. **Engineering Structures**, v. 130, p. 180-197, 2017.

MORRISON, D. F. **Multivariate Statistical Methods**, 4 ed. Philadelphia. Duxbury Advanced Series, 2005. 498p.

MUTIHAC, L; MUTIHAC, R. Mining in chemometrics. **Analytica Chimica**, v. 612, p. 1-18, 2008.

NADADOOR, V. R.; SIEGLER, H. D. L. H.; SHAH, S.; MCCAFFREY, W. C.; BEN-ZVI, A. B. Online sensor monitoring a microalgal bioreactor system using support vector regression. **Chemometrics and Intelligent Laboratory Systems**, v.110, p. 38-48, 2012.

NAES, T.; ISAKSSON, T.; FEARN, T.; DAVIES, T. Partial Least Squares. In: **Multivariate calibration and classification**. Chichester. NIR Publications, 2002.

NAES, T.; MARTENS, H. Multivariate calibration II. Chemometric methods. **Trends in analytical chemistry**, v. 3, p. 266-271, 1984.

NI, Y.; MEI, M.; KOKOT, S. Analysis of complex, processed substances with use of NIR spectroscopy and chemometrics: Classification and prediction of properties – The potato crisps example. **Chemometrics and Intelligent Laboratory Systems**, v.105, p. 147-156, 2011.

NICODEM, D. E.; FERNANDES, M. C. Z.; GUEDES, C. L. B.; CORREA, R. J. Photochemical processes and environmental impact of petroleum spills. **Biogeochemistry**, v. 39, p. 121-138, 1997.

NOORI, R.; KARBASSI, A. R.; MOGADDAMNIA, A.; HAN, D.; ASHTIANI, M. H. Z.; FAROKHNIA, A.; GOUSHEH, M. G. Assessment of input variables determination on the SVM model performance using PCA, gamma test, and forward selection techniques for monthly stream prediction. **Journal of Hydrology**, v.401, p.177-189, 2011.

OLAFSSON, S.; LI, X.; WU, S. Operations research and data mining. **European Journal of Operational Research**, v.187, p. 1429 – 1448, 2008.

OTERO, R. L. S.; GALVÃO, R. K. H.; ARAÚJO, M. C. U.; CAVALHEIRO, E. T. G. Thermogravimetric determination of L-ascorbic acid in non-effervescent by the successive projections algorithm. **Thermochimica Acta**, v.526, p. 200-204, 2011.

PAIVA, A. P. **Metodologia de superfície de resposta e análise de componentes principais em otimização de processos de manufatura com múltiplas respostas correlacionadas**. 2006. 129 f. Tese (Doutorado em Engenharia Mecânica) – Instituto de Engenharia Mecânica, Universidade Federal de Itajubá, UNIFEI, 2006.

PHADKE, M. In: **Quality Engineering using Robust Design**. Prentice Hall, 1989.

PAL, A.; MAITI, J. Development of a hybrid methodology for dimensionality reduction in Mahalanobis–Taguchi system using Mahalanobis distance and binary particle swarm optimization. **Expert Systems with Applications**, v. 37, p. 1286-1293, 2010.

PAN. Q.; DIAS, D. An efficient reliability method combining adaptive Support Vector Machine and Monte Carlo Simulation. **Structural Safety**, v.67, p. 85-95, 2017.

PASQUINI, C. Near Infrared Spectroscopy: Fundamentals, Practical Aspects and Analytical Applications. **Journal of Chemical Brazilian Society** , v. 14, p. 198-219, 2003.

PASQUINI, C.; BUENO, A. F. Characterization of petroleum using near-infrared spectroscopy: Quantitative modeling for the true boiling point curve and specific gravity. **Fuel**, v. 86, p. 1927-1934, 2007.

PASTI, F. E. L.; CENTNER, V.; WALCZAK, B.; DESPAGNE, F.; RIMBAUD, D. J.; NOORD, O. E.; MASSART, D. L. A comparison of multivariate calibration techniques applied to experimental NIR data sets Part II. Predictive ability under extrapolation conditions. **Chemometrics and Intelligent Laboratory Systems**, v. 58, p. 195-211, 2001.

PENG, J.; LI, L. Support vector regression in sum space for multivariate calibration. **Chemometrics and Intelligent Laboratory Systems**, v.130, p. 14-19, 2014.

PESSOA, C. M.; RANZAN, C.; TRIERWEILER, L. F.; TRIERWEILER, J. O. Development of ant colony optimization (ACO) algorithms based on statistical analysis and hypothesis testing for variable selection. **IFAC-Papers On Line**, v.48, p. 900-905, 2015.

PUGNA, A.; NEGREA, R.; MICLEA, S. Using six sigma methodology to improve the assembly process in an automotive company. **Procedia Social and Behavioral Sciences**, v.221, p. 308-316, 2016.

PRAKASH, J.; MISHRA, A. K. Introduction of a novel fluorescence based method for the composition estimation of petroleum fuel-biofuel blends. **Fuel**, v.108, p. 351-355, 2013.

RANZAN, C. **Desenvolvimento de Modelos Quimiométricos utilizando o Algoritmo de Otimização Colônia de Formigas**. 2014. 75 f. Tese (Doutorado em Engenharia) – Departamento de Engenharia Química, Universidade Federal do Rio Grande do Sul, UFRS, 2014.

RESÉNDIZ, E.; RULL-FLORES, C. Mahalanobis-Taguchi system applied to variable selection in automotive pedals components using Gompertz binary particle swarm optimization. **Expert Systems Applications**, v.40, p. 2361-2365, 2013.

RONCHETTI, E.; FIELD, C.; BLANCHARD, W. Robust linear model selection by crossvalidation. **Journal of the American Statistical Association**, v. 92, p. 1017-10231, 1997.

SAMMANI, M.; CLAVIJO, S.; PORTUGAL, L.; SUÁREZ, R.; SEDDIK, H.; CERDÀ, V. Use of multiresponse statistical technics to optimize the separation of diosmin, hespediridin, dismetin and different pharmaceutical preparations by high performance liquid cromatography with UV-DAD. **Talanta**, v. 167, p. 695-702, 2017.

SHENG, G. X.; JING-SHUN, D.; JIA-FU, W.; WEI, C. Anomaly intrusion detection based on PLS feature extraction and core vector machine. **Knowledge-Based Systems**, v.40, p. 1-6, 2013.

SILVA, A. P. D. Optimization approaches to supervised classifications. **European Journal of Operational Research**, v.261, p. 772-788, 2017

SILVA, M. P. F.; BRITO, L. R.; HONORATO, F. A.; PAIM, A. P. S.; PASQUINI, C.; PIMENTEL, M. F. Classification of gasoline as with or without dispersant and detergent additives using infrared spectroscopy and multivariate calibration. **Fuel**, v. 116, p. 151-157, 2014.

SARTAKHTI, J. S.; ZANGOOEI, M. H.; MAZAFARI, K. Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA). **Expert Systems with Applications**, v.108, p. 570-5791, 2012.

SHI, W.; SHANG, J.; LIU, Z.; ZUO, X. Optimal design of the auto parts supply chain for JIT operations: Sequential bifurcation factor screening and multi-response surface methodology. **European Journal of Operational Research**, v.236, p. 664-676, 2014.

SKOOG, D. A.; HOLLER, F. J.; CROUCH, S. R. **Princípios de Análise Instrumental**, 6 ed. Porto Alegre: Editora Bookman, 2007. 1055 p.

SOARES, S. F. C.; GOMES, A. A.; FILHO, A. R. G.; ARAUJO, M. C. U.; GALVÃO, R. K. H. The successive projections algorithm. **Trends in Analytical Chemistry**, v.42, p. 84-98, 2013.

SUN, X.; ZIMMERMANN, C. M.; JACKSON, G. P.; BUNKER, C. E. Classification of jet fuels by fuzzy rule-building expert systems applied to three-way data by fast gas chromatography-fast scanning quadrupole ion trap mass spectrometry. **Talanta**, v.83, p. 1260-1268, 2011.

THISSEN, U.; PEPERS, M.; ÜSTÜN, B.; MELSEN, E. J.; BUYDENS, L. M. C. Comparison support vector machines to PLS spectral regression applications. **Chemometrics Intelligent Laboratory Systems**, v.73, p. 169-179, 2004.

TOMREN, A. L.; BARTH, T. Comparison of partial least squares calibration models of viscosity, acid number of asphaltene content in petroleum, base on GC and IR data. **Fuel**, v. 120, p. 8-21, 2014.

TONG, P.; DU, Y.; ZHENG, K.; WU, T.; WANG, J. Improvement of NIR model by fractional order Savitzky-Golay derivation (FOSGD) coupled with wavelength selection. **Chemometrics and Intelligent Laboratory Systems**, v. 143, p. 50-48, 2015.

VAPNIK, V. N. The Naure of Statistical Learning Theory, Springer, New York, 1995.

WANG, S.; HUANG, G. A multi-level Taguchi-factorial two-stage stochastic programming approach for characterization of parameter uncertainties and their interactions: An application to water resources management. **European Journal of Operational Research**, v. 240, p. 572 – 581, 2015.

WANG, Y.; ZHAO, X.; WANG, B. LS-SVM and Monte Carlo methods based reliability analysis for settlement of soft clayey foundation. **Journal of Rock Mechanics and Geotechnical Engineering**, v.5, p. 312-317, 2013.

WANG, X.; WEN, J.; ZHANG, Y.; WANG, Y. Real state price forecasting based on SVM optimizes by PSO. **Optik**, v.125, p. 1439-1443, 2014.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 58, p. 109-130, 2001.

WU, T.; HSU, M. F. Credit risk assessment and decision making by a fusion approach. **Knowledge-Based Systems**, v.35, p.102-110, 2012.

XU, Z.; BUNKER, C. E.; HARRINGTON, P. B. Classification of jet fuel properties by near-infrared spectroscopy using fuzzy rule-building expert systems and support vector machines. **Applied Spectroscopy**, v.64, p. 1251-1258, 2010.

XU, S.; LU, B.; BALDEA, M.; EDGAR, T. F.; NIXON, M. Na improvement variable selection method for support vector regression in NIR spectral modeling. **Journal of Process Control**, v. Article in Press, 2017.

YEH, C. C.; CHI, D. J.; HSU, M. F. A hybrid approach of DEA, rough and support vector machines for business failure prediction. **Expert Systems with Applications**, v.37, p. 1535-1541, 2010.

YI-HUI, C. Using combined AHP and PLS path modelling on blog site evaluation in Taiwan. **Computers in Human Behavior**, v.29, p. 1325-1333, 2013.

YIP, W. L.; SOOSAINATHER, T. C.; DURSTAD, K.; SANDE, S. A. Classification of structurally related commercial contrast media by near infrared spectroscopy. **Journal of Pharmaceutical and Biomedical Analysis**, v.90, p. 148-160, 2014.

ZHANG, W.; YUAN, W.; ZHANG, X.; CORONADO, M. Prediction the dynamic and kinematic viscosities of biodiesel-diesel blends using mid and near-infrared spectroscopy. **Applied Energy**, v.98, p. 122-127, 2012.

ZHANG, M.; ZHANG, S.; IQBAL, J. Key wavelengths selection from near infrared spectra using Monte Carlo sampling-recursive partial least squares. **Chemometrics and Intelligent Laboratory Systems**, v.128, p. 17-24, 2013.

ZOLGHARNEIN, J.; ASANJRANI, N.; BAGTASH, M.; AZIMI, G. Multiple-response optimization using Taguchi design and principle component analysis for removing binary mixture of alizarin red and alizarin yellow from aqueous solution by nano γ -alumina. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v.126, p. 291-300, 2014.

ZAMORA, P. P.; PONCE, L. C.; NAGATA, N.; POPPI, R. J. Alternativas quimiométricas para a resolução de problemas analíticos clássicos. Determinação espectrofotométrica de misturas de zircônio e háfnio. **Química Nova**, v. 20, p. 469-474, 1997.