UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

# METODOLOGIAS PARA SELEÇÃO DE VARIÁVEIS EXPLICATIVAS E DETECÇÃO DE INCONFORMIDADES DE PREDIÇÃO APLICADAS À ESPECTROSCOPIA POR FLUORESCÊNCIA

TESE DE DOUTORADO

*Lucas Ranzan*

**Porto Alegre**

**2021**

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

DEPARTAMENTO DE ENGENHARIA QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

# METODOLOGIAS PARA SELEÇÃO DE VARIÁVEIS EXPLICATIVAS E DETECÇÃO DE INCONFORMIDADES DE PREDIÇÃO APLICADAS À ESPECTROSCOPIA POR FLUORESCÊNCIA

*Lucas Ranzan*

Tese de Doutorado apresentada como requisito parcial para obtenção do título de Doutor no Programa de Pós-Graduação em Engenharia Química da UFRGS

*Área de concentração*: Pesquisa e Desenvolvimento de Processos

*Linha de Pesquisa*: Projeto, Simulação, Modelagem, Controle e Otimização de Processos Químicos e Bioprocessos.

**Orientadores:**
**Prof. Dr. Jorge Otávio Trierweiler**
**Profª.Drª. Luciane Ferreira Trierweiler**

**Porto Alegre**

**2021**

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

ESCOLA DE ENGENHARIA

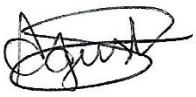
DEPARTAMENTO DE ENGENHARIA QUÍMICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

A Comissão Examinadora, abaixo assinada, aprova a Proposta de Pesquisa *Metodologias para Seleção de Variáveis Explicativas e Detecção de Inconformidades de Predição Aplicadas à Espectroscopia por Fluorescência,* elaborada por Lucas Ranzan, como requisito parcial para obtenção do Grau de Doutor em Engenharia.

Comissão Examinadora:

Prof. Dr. Adilson Ben da Costa



Prof. Dra. Caroline Borges Agustini

Prof. Dr. Michel J. Anzanello

# Resumo

A capacidade de predizer eventos futuros a partir de conhecimentos históricos é a base para a modelagem preditiva. Criar um modelo capaz de quantificar variáveis de interesse, classificar ocorrências ou prever comportamentos, acompanham a evolução dos algoritmos modernos de aprendizado de máquina. Na indústria de transformação, muitas das informações mais relevantes para o controle de processos ainda são adquiridas unicamente através de técnicas laboratoriais, que são custosas, destrutivas e morosas (como, por exemplo, concentração molecular de espécies de interesse, pureza de fármacos, lubricidade de óleos, teor de proteína em alimentos, etc.). Um possível caminho para automação destes sistemas é o estudo de novos sensores capazes de capturar uma informação auxiliar de fácil obtenção, que possa ser transformada matematicamente nas saídas de interesse. Surge então a aspiração por estudos que combinam a escolha de sensores adequados com metodologias capazes de extrair de maneira eficiente a informação útil contida nestes dados. Neste trabalho são apresentadas metodologias baseadas em diferentes estratégias para seleção de variáveis explicativas e otimização de modelos empíricos. Ainda, é proposta uma metodologia para qualificação de inconformidades em novas leituras utilizando redes neurais. É apresentada a metodologia AnTSbe, um algoritmo híbrido baseado nas meta-heurísticas Colônia de Formigas (ACO) e Busca Tabu (TS), desenvolvido para otimizar a seleção de variáveis de entrada em problemas combinatórios complexos. A hibridização das meta-heurísticas visa evitar a estagnação precoce e a ciclagem de subgrupos, comuns nessas metodologias. O algoritmo também introduz o uso da expansão polinomial e combinatória das variáveis de entrada, em um esforço para incrementar o poder preditivo dos modelos. Como estudo de caso, espectroscopia por fluorescência é utilizada para predizer concentração de enxofre em diesel combustível. Os modelos preditivos ajustados foram superiores a outras técnicas descritas na literatura, com erros absolutos percentuais médios de predição menores que 4%. As adaptações propostas se mostraram eficientes, quando comparadas a pesquisas prévias com a mesma base de dados. Uma adaptação é proposta ao algoritmo AnTSbe, focada para dados de fluorescência, com o conceito de Delta Pair. Uma nova camada de otimização é introduzida no algoritmo a fim de selecionar um par Excitação/Emissão que serve como regulador do meio, tendo sua intensidade de fluorescência decrescida de todos outros os pontos do espectro. Neste estudo, são acompanhados três processos distintos de envelhecimento de cachaça, com o intuito de predizer a concentração de fenólicos na bebida ao longo do tempo, com base em dados fluorescência. A adaptação Delta Pair se mostrou especialmente funcional quando combinada com expansão de bases e para predição de cachaças envelhecidas comerciais, que não participaram da etapa de calibração dos modelos. A seguir, matrizes excitação – emissão de fluorescência captadas *in situ* em fermentações com *S. cerevisiae* foram utilizadas para calibrar uma rede neural convolucional residual, como intuito de predizer glicose, etanol e biomassa no meio biológico. Em paralelo, foi desenvolvida uma metodologia baseada em redes neurais do tipo *autoencoder* (AE), capazes de corretamente reconstruir os espectros originais. A metodologia utiliza o erro de reconstrução da rede AE treinada para triagem não supervisionada de novos espectros, conseguindo identificar espectros com inconformidades, e qualificar a confiança que se pode atribuir a um novo dado, baseado na magnitude deste erro. Por fim, a metodologia AnTSbe é utilizada para predizer impurezas

nas correntes de uma unidade de separação de propano/propeno, expandindo o uso da metodologia para casos da indústria petroquímica com base em dados simulados de processo (e não de fluorescência). A metodologia se mostrou capaz de corretamente predizer os perfis de concentração das três colunas de separação do processo com erros absolutos percentuais médios inferiores a 5%, com foco especial para quantificação dos contaminantes em cada corrente, que precisam ser mantidos sob controle para garantir a lucratividade da operação. Os artigos desenvolvidos demonstram, inclusive na ordem apresentada, o sucesso das metodologias propostas em aprofundar a seleção de variáveis significativas e otimização de modelos empíricos preditivos. A sucessão dos casos estudados parte do desenvolvimento do algoritmo estocástico base, segue para a busca de um reforço na capacidade de generalização dos modelos otimizados baseados em espectroscopia por fluorescência, apresenta uma técnica para qualificação de novas amostras e conclui com o uso dos algoritmos desenvolvidos em um caso industrial.

# Abstract

The ability to predict future events from historical observations is the basis for predictive modeling. Creating a model capable of quantifying variables of interest, classifying occurrences or predicting behavior, follows the evolution of modern machine learning algorithms. In the manufacturing industry, much of the most relevant information for process control is still acquired only through laboratory techniques, which are costly, destructive and time-consuming (such as, for example, molecular concentration of species, purity of drugs, lubricity of oils, protein content in food, etc.). A possible way to automate these systems is the study of new sensors capable of capturing auxiliary information of easy application, which can be mathematically transformed in the outputs of interest. This is the aspiration for studies that combine the choice of skilled sensors with methodologies capable of efficiently extracting the useful information contained in the data. In this work we propose methodologies based on different machine learning methods for the optimization of empirical models. AnTSbe methodology is presented, a hybrid algorithm based on Ant Colony (ACO) and Tabu Search (TS) metaheuristics, developed to optimize the selection of input variables in complex combinatorial problems. The hybridization of metaheuristics aims to avoid early stagnation and cycling of subgroups, common in these methodologies. The algorithm also introduces the use of polynomial and combinatorial expansion of the input variables, in an effort to increase the predictive power of the models. As a case study, fluorescence spectroscopy is used to predict sulfur concentration in diesel fuel. The adjusted predictive models were superior to other techniques from literature, with mean absolute percentage errors of prediction smaller than 4%. The proposed adaptations were efficient, when compared to previous researches with the same database. An adaptation is proposed to the AnTSbe algorithm, focused on fluorescence data, with the concept of DeltaPair. A new optimization layer is introduced in the algorithm in order to select an Excitation/Emission pair that serves as a medium regulator, having its fluorescence intensity decreased from all other points in the spectrum. In this study, three distinct cachaça aging processes are followed, in order to predict the concentration of phenolics in the spirit over time, based on fluorescence data. The DeltaPair adaptation is especially functional when combined with base expansion and for the prediction of aged commercial cachaças, which does not participate in the calibration stage of the models. Following, fluorescence excitation - emission matrices, collected *in situ* in fermentations with *S. cerevisiae,* were used to calibrate a residual convolutional neural network, in order to predict glucose, ethanol and biomass in the biological environment. In parallel, a methodology based on autoencoder-type neural networks (AE) was developed, capable of correctly reconstructing the original spectra. The methodology uses the trained AE reconstruction error for unsupervised screening of new spectra, managing to identify abnormal spectra, and to qualify the confidence that can be attributed to a new data, based on the magnitude of this error. Despite the focus on fluorescence spectroscopy data, most of the methodologies were designed to be of general use, whatever the data source, with little or no modification. Finally, the AnTSbe methodology is used to predict impurities in

the streams of a propane/propylene splitter unit, expanding the use of the methodology for cases in the petrochemical industry based on simulated process data (and not fluorescence). The methodology proved to be capable of correctly predicting the concentration profiles of the three process' separation columns with mean absolute percentage errors below 5%, with a special focus on quantifying the contaminants in each stream, which need to be kept under control to ensure profitability of the operation. The articles developed demonstrate, in the order presented, the success of the proposed methodologies in deepening the selection of significant variables and the optimization of predictive empirical models. The succession of the studied cases starts from the development of the base stochastic algorithm, goes on to seek a reinforcement in the generalizability of the optimized models based on fluorescence spectroscopy, presents a technique for qualifying new samples and concludes with the use of the algorithms developed in an industrial case.

*“Tanto a estrada quanto a história têm sido longas, você não concordaria? A viagem tem sido longa e o custo tem sido alto...mas nunca uma coisa grande foi alcançada com facilidade. Uma longa história, como uma Torre alta, tem que ser construída pedra por pedra”.*

*(Stephen King – A Torre Negra)*

# Agradecimentos

Inicio estes agradecimentos dirigindo minhas palavras as pessoas que, sem dúvida nenhuma, foram as responsáveis por eu poder estar aqui escrevendo este texto: minha família. Mãe e pai, obrigado por me darem todas as condições de suporte, amor e carinho para seguir minhas escolhas e meu próprio caminho. Mano, muito obrigado por estar sempre do meu lado, principalmente nas horas mais difíceis, quando a esperança e a força de vontade falhavam. Vocês sempre foram e sempre serão meus maiores modelos, não só por me mostrarem, cada um do seu jeito, a importância da educação, mas por serem a representação física e palpável do que eu considero o amor incondicional de uma família.

Agradeço imensamente aos meus orientadores Jorge Otávio Trierweiler e Luciane Ferreira Trierweiler, sem os quais esta tese não existiria. Vocês não só foram importantes para todo o desenvolvimento conceitual e a aplicação prática das ideias que desenvolvemos neste trabalho, como tem sido a base sólida sobre a qual construí grande parte da minha vida acadêmica. Obrigado por todo suporte ao longo desses vários anos, pelos ensinamentos, por apoiarem as minhas ideias e por todas as oportunidades que me proporcionaram. Agradeço ao Prof. Dr. Bernd Hitzmann e aos colegas que me acolheram em seu grupo de pesquisa em Hohenheim.

Sem dúvida, não posso deixar de agradecer aos amigos e amigas que trilharam ao meu lado essa longa caminhada. Muito obrigado aos meus irmãos e agregados da G.M., a minha vida não teria metade da graça sem vocês nela. Obrigado Karen e Twin, por sempre serem meus portos-seguros. Sei que vocês nunca duvidaram, por nenhum segundo sequer, que eu chegaria até aqui. Obrigado ao Ka-tet Chambourcy, Presuntinho, Carol, Ju, Yana, Judel, Pistolinha e aos amigos que a eng. química me proporcionou. Vocês foram e continuam sendo parte especial de mim. Agradeço a Marina por dividir comigo por muitos anos a carga extenuante de produzir esse trabalho. Obrigado Elisa por estar ao meu lado no fim desta jornada, pelo conforto e por todo apoio nessa nova fase que se inicia na minha vida. Obrigado aos amigos que me acolheram e me ajudaram durante a minha estadia na Alemanha. Nenhuma jornada é construída sozinho, e eu tenho sorte em poder ter dividido essa com vocês. Agradeço aos colegas e amigos que fiz no DEQUI, em especial ao Pedro e aos demais que participaram ativamente como colaboradores das minhas pesquisas, e aos alunos que tive que a honra de orientar.

Por fim, agradeço a todas as pessoas que tiveram parte neste trabalho e na minha vida nestes anos. Muito Obrigado.

**SUMÁRIO**

**LISTA DE FIGURAS**

**LISTA DE TABELAS**

# Capítulo 1 – **Introdução**

## 1.1 Motivação

Na sociedade do século XXI, o uso de modelagem preditiva permeia todas as áreas do conhecimento. Criar um modelo com a capacidade de quantificar variáveis de interesse, classificar ocorrências ou até mesmo prever resultados futuros através do estudo de dados são as bases de muitas das evoluções tecnológicas das últimas décadas. Desde detectar câncer (EINIPOUR; CORRESPONDING, 2011; SUN et al., 2019), prever comportamento de mídias sociais (MORO; RITA; VALA, 2016), até regular produções industriais (BELTRAMO; KLOCKE; HITZMANN, 2019), novos algoritmos de *machine learning* são desenvolvidos constantemente para predizer informações que seriam muito custosas, invasivas, ou impossíveis de medir diretamente.

Em uma escala global, os sistemas industriais e de análise de comportamento social se tornam cada vez mais competitivos. A chamada Quarta Revolução Industrial se baseia no conceito de retirar o componente humano das tomadas de decisões que possam ser substituídas por inteligência artificial. Além disso, a interconexão de máquinas e sistemas, o desenvolvimento de sensores inteligentes e o rastreio e armazenamento de dados tornam possíveis a automação de diversos processos que seriam impraticáveis em décadas passadas (LU, 2017). Em grandes áreas econômicas como a indústria do petróleo e de alimentos, os processos avançados de controle possibilitam uma vantagem econômica inegável, e é provável que empresas que não se adaptarem a esta evolução sejam, aos poucos, excluídas do mercado.

No cenário atual, muitas das informações mais relevantes para o controle de processos da indústria de transformação (como concentração de espécies de interesse, ponto de ebulição de solventes, viscosidade, lubricidade de óleos, teor de proteínas em alimentos, concentração de biomassa em sistemas biológicos, pureza de fármacos, etc.) ainda são adquiridas unicamente por meio de testes laboratoriais, que são custosos, destrutivos, demorados e requerem pessoal e instalações especializadas. Um possível caminho para a automação destes sistemas pode se dar na forma de sensores que capturem informações auxiliares de fácil aplicação, que por sua vez possam ser transformadas em inferidores

virtuais das variáveis de difícil aferição direta. A abordagem de criar um modelo que relacione os dados de entrada com as saídas de interesse, sem a necessária compreensão dos mecanismos físicos envolvidos, é denominada modelagem empírica.

A evolução tecnológica dos sistemas de coleta e armazenamento de dados faz com que diariamente uma quantidade considerável de informação seja capturada (por exemplo, na forma de historiadores de processo e imagens em alta resolução). Porém, tão importante quanto a coleta dos dados, são os mecanismos para extrair a informação útil dos mesmos. Uma grande quantidade de variáveis geralmente está associada a um alto nível de ruído, de colinearidade e de possíveis variáveis irrelevantes ou redundantes (TANG; ALELYANI; LIU, 2014). Para resolver esses problemas, são empregadas técnicas de extração e seleção de variáveis, conseguindo descrever o sistema de maneira eficiente em um espaço reduzido.

Assim, ao ansiar por sistemas de produção mais automatizados e eficientes, surge a necessidade de estudos que combinam a escolha de sensores hábeis com o desenvolvimento de metodologias de seleção de variáveis capazes de otimizar o ajuste de modelos empíricos preditivos para as saídas de interesse.

Neste contexto, a espectroscopia por fluorescência apresenta diversas características que propiciam sua aplicação como sensor útil em sistemas químicos e biológicos. Diversas moléculas de interesse nestes processos são fluoróforos naturais, como proteínas, hidrocarbonetos poliaromáticos, compostos heterocíclicos e alifáticos altamente insaturados. A aquisição dos espectros é rápida e não demanda manipulação obrigatória das amostras, podendo ser utilizada diretamente em linha para acompanhamento do sistema em tempo real. Por se tratar de uma técnica óptica, seu uso é não-invasivo e não-destrutivo, podendo ser interfaceada no sistema através de janelas de quartzo, o que a torna viável para ambientes com risco de explosão/incêndio ou cultivo biológico. De um único espectro pode-se predizer uma gama de informações sobre características e constituintes do meio. Em especial, é uma técnica sensível, capaz de identificar espécies orgânicas e inorgânicas em quantidades traço e permite, muitas vezes, determinação qualitativa e quantitativa de substâncias contidas na amostra (SKOOG; HOLLER; CROUCH, 2009).

Em contraponto às vantagens apresentadas, a otimização de modelos empíricos baseados em dados de fluorescência não é de maneira alguma trivial. Dependendo da resolução, as matrizes excitação-emissão de fluorescência (EEM) podem conter entre centenas e milhares de pares excitação/emissão, em uma distribuição altamente colinear e contendo informações que podem não ter relação com a variável ou propriedade de interesse a qual se pretende predizer. Por esses motivos, as estratégias de seleção de variáveis propostas precisam ter um bom equilíbrio entre os mecanismos de prospecção e exploração, conseguindo vagar entre a enormidade de combinações possíveis, mas evoluindo em direção aos modelos mais representativos. Neste âmbito, metodologias baseadas em meta-heurísticas, como as do tipo Colônia de Formigas (ACO), são reconhecidas na literatura como eficientes para lidar com problemas complexos de análise combinatória, ao exemplo da resolução de problemas do tipo Caixeiro Viajante, no qual foram primeiramente apresentadas (DORIGO; GAMBARDELLA, 1997). Os algoritmos ACO já foram aplicados previamente em nosso grupo de pesquisa para otimizar com sucesso

modelos baseados tanto em dados de espectroscopia por fluorescência, quanto para outras técnicas espectrométricas (RANZAN et al., 2011, 2014, 2015, 2017). Porém, a grande capacidade adaptativa destes algoritmos abre possibilidades para diversas melhorias que buscam resolver desafios inerentes aos métodos estocásticos iterativos, como estagnação precoce, ciclagem de subgrupos e problemas de generalização. Dentre as adaptações sugeridas, a hibridização com outras técnicas de otimização pode ser vista como caminho natural para solução dos problemas supracitados (BHATTACHARYYA, 2018). A proposta de evolução das meta-heurísticas visa, assim, não apenas ajustar melhores modelos preditivos, mas garantir uma maior eficiência dos algoritmos, com resultados mais consistentes e maior aplicabilidade a diferentes casos de estudo.

Em uma outra abordagem ao ajuste de modelos caixa-preta, é inegável que as técnicas de aprendizado de máquina mais utilizadas e pesquisadas na última década se referem ao uso de Redes Neurais Artificiais (ANN). Modelos baseados nas mais diversas configurações de ANN são o estado-da-arte em múltiplos campos, especialmente no que diz respeito ao reconhecimento de padrões (ABIODUN et al., 2018; JES et al., 2019). Majoritariamente aplicadas para tratamento e classificação de imagens, as Redes Neurais Convolucionais (CNN) são, possivelmente, as mais populares e efetivas arquiteturas atuais de aprendizado profundo (*deep learning*) (RUIZ-DEL-SOLAR; LONCOMILLA; SOTO, 2018). Seu uso, porém, ainda é incipiente na construção de modelos quimiométricos. Isso pode ser ligado ao fato de que tais redes necessitam de uma vasta disponibilidade de dados de segunda ordem para seu treinamento, que são amplamente acessíveis no caso de imagens, mas muitas vezes de difícil aquisição no contexto de processos químicos/biológicos. No escopo deste trabalho, é promissor o emprego de CNN no ajuste de modelos preditivos, utilizando como base de dados matrizes excitação – emissão de fluorescência coletadas continuamente *in situ*. O acoplamento de um espectrofluorímetro a um sistema possibilita um influxo contínuo de EEMs, que alimentam a rede neural para qualificação e quantificação de variáveis de interesse do meio, de maneira a reduzir a dependência de análises laboratoriais ou proporcionar informações *online* para ações de controle.

Uma das limitações dos modelos empíricos é sua baixa capacidade de generalização e extrapolação. Diferentemente de modelos baseados em conceitos fenomenológicos, os modelos empíricos são totalmente dependentes dos dados de treinamento e teste, e só se pode confiar em sua capacidade preditiva dentro de um contexto semelhante ao destas amostras, como, por exemplo, dentro de uma mesma faixa de concentração. Assim, é importante um acompanhamento periódico da qualidade das predições do modelo com metodologias *off-line*, como testes laboratoriais. Porém, o tempo morto associado a estas medidas *off-line* é justamente o que se espera superar com o uso dos inferidos virtuais, e é possível que o sistema passe longos períodos sendo monitorado virtualmente antes que se obtenha confirmação externa de normalidade. Neste meio tempo é concebível que o sistema seja alimentado por entradas que divirjam significativamente do conjunto de calibração. Tomar ações baseadas em predições indignas de confiança pode acarretar inúmeros prejuízos ao sistema. Assim, surge a necessidade de uma metodologia para se qualificar de forma não-supervisionada dados de entrada, com a intenção de avaliar se os mesmos se enquadram nas características dos conjuntos previamente utilizados para calibração dos preditores, e se suas predições devem ser confiadas.

Nesta conjectura, as redes neurais do tipo *AutoEncoder* (AE) (CHARTE et al., 2018) apresentam potencial para triagem e qualificação de novos dados de entrada. As redes AE utilizam uma estrutura simétrica para primeiramente transformar as entradas em um novo

espaço, chamado *encodado*. Nele, características dos dados são explicitadas, em uma forma similar ao espaço criado na Análise de Componentes Principais (PCA), mas com a capacidade de captura de informação não linear (ALMOTIRI; ELLEITHY; ELLEITHY, 2017; JES et al., 2019). O espaço encodado é então *decodificado*, buscando reproduzir as entradas originais. Assim, os parâmetros da rede são otimizados para diminuir o erro de reconstrução entre as saídas e as entradas (que almejam serem as mesmas). Uma rede AE ajustada é capaz de quantificar quanto uma nova entrada é semelhante aquelas com as quais ela foi treinada, com base no seu erro de reconstrução. Aplicações de triagem e reconhecimento de dados anormais podem ser vistas como no caso de identificação de fraudes em cartão de credito (MISRA et al., 2020; ZOU; ZHANG; JIANG, 2019), na detecção de anomalias em dados de satélites (SAKURADA; YAIRI, 2014) e no diagnóstico computacional de diversas patologias (AMARBAYASGALAN; JARGALSAIKHAN; RYU, 2018). É de interesse avaliar a capacidade de uma rede AE treinada com dados de espectroscopia por fluorescência em detectar falhas ou desvios em novas leituras espectrais, e quantificar a confiabilidade nestes dados.

Com o objetivo de aprofundar os conhecimentos em seleção de variáveis e otimização de modelos empíricos preditivos, neste estudo adaptações e metodologias serão propostas, como a hibridização de algoritmos meta-heurísticos e modificações nos mecanismos de exploração e tratamento dos dados de entrada, assim como o desenvolvimento de técnicas baseadas em redes neurais para ajuste de modelos preditivos e reconhecimento de padrões. As metodologias apresentadas serão então avaliadas utilizando, majoritariamente, dados de espectroscopia por fluorescência para o desenvolvimento de técnicas alternativas às convencionais para predição de variáveis de interesse em diferentes estudos de caso.

## 1.2 Objetivos e Estruturação do Trabalho

O objetivo principal deste trabalho trata sobre o desenvolvimento e evolução de metodologias baseadas em meta-heurísticas para seleção de variáveis e otimização de modelos empíricos, incluindo todas as facetas de tratamento de dados necessários para tal. Além disso, serão propostas metodologias baseadas em redes neurais para predição de variáveis de processo e desenvolvimento de uma rotina de triagem capaz de qualificar e detectar inconformidades na base de dados futuros. Unindo as metodologias se encontra o uso de matrizes de dados obtidos através de espectroscopia por fluorescência bidimensional nos estudos de caso.

Os objetivos específicos (**$O_n$**) são apresentados na sequência:

- **$O_1$** Propor novas adaptações para algoritmos de seleção de variáveis baseados em meta-heurísticas.

- **$O_2$** Desenvolver uma metodologia capaz de aprimorar a generalização de modelos empíricos baseados em dados de espectroscopia por fluorescência.

- **$O_3$** Desenvolver uma metodologia baseada em redes neurais para triagem e identificação de dados anômalos.

- **$O_4$** Avaliar a aplicação de redes neurais convolucionais para predição de variáveis de processos com coleta *online* de espectros.

- **$O_5$** Avaliar o uso de espectroscopia por fluorescência para predição de variáveis de interesse que atualmente são quantificadas majoritariamente por testes laboratoriais.

- **$O_6$** Expandir a aplicação das metodologias desenvolvidas em estudos de caso da indústria petroquímica que não envolvam dados de fluorescência.

No Capítulo 2 será apresentada uma revisão bibliográfica, contemplando a fundamentação teórica dos assuntos mais relevantes contidos neste trabalho e pertinentes ao seu entendimento.

O Capítulo 3 descreve minunciosamente o desenvolvimento do algoritmo de otimização de modelos empíricos AnTSbe. O artigo descreve as etapas iniciais de tratamento e segmentação dos dados, a metodologia proposta de seleção de variáveis baseada em uma hibridização entre meta-heurísticas Colônia de Formigas e Busca Tabu, e introduz o conceito de expansão de variáveis de entrada. A metodologia é aplicada a um estudo de caso, utilizando espectroscopia por fluorescência para predizer concentração de enxofre em amostras de diesel combustível.

O Capítulo 4 apresenta um estudo sobre caracterização de perfis e predição de compostos fenólicos totais em processos de envelhecimento de cachaça em barris de amburana. Três processos de envelhecimento foram realizados em nosso departamento, em uma pesquisa que perdurou por mais de três anos. Ao longo de cada envelhecimento, amostras eram retiradas e seu espectro de fluorescência coletado. A concentração total de fenólicos em cada amostra também foi quantificada por métodos tradicionais. A metodologia proposta para predição da concentração de fenólicos é uma adaptação do algoritmo AnTSbe, e tem como principal objetivo a introdução com uma nova camada externa de seleção de variáveis, criando o conceito de par delta, um par excitação/emissão que é utilizado como referência interna em cada espectro. Os resultados deste trabalho corroboram fortemente com as contribuições do Capítulo 3.

O Capítulo 5 trata sobre o desenvolvimento de uma metodologia que utiliza o erro de reconstrução de uma rede neural do tipo autoencoder como ferramenta de triagem para amostras anômalas. No estudo, espectros por fluorescência são coletados *in situ* durante três processos fermentativos utilizando a levedura *Saccharomyces cerevisiae*. Tais dados são alimentados em dois processos paralelos: o primeiro faz uso de redes neurais residuais para ajustar um modelo de predição para etanol, glicose e biomassa do meio. O segundo utiliza os dados para criar uma rede autoencoder capaz de corretamente representar o conjunto de espectros com erros de reconstrução aceitáveis. Para simular novas amostras, alguns pontos do conjunto de dados original são dopados com erros, e o novo erro de reconstrução da rede autoencoder treinada é utilizado para qualificar as amostras anômalas. Este estudo foi iniciado sob a tutela do Prof. Dr. Bernd Hitzmann, em parceria com a Universidade de Hohenheim – Alemanha, durante o período de doutorado-sanduíche do autor.

O Capítulo 6 relata a expansão da aplicação da metodologia AnTSbe para o desenvolvimento de sensores virtuais para predição de contaminantes em uma Unidade de Separação Propano/Propeno, com o uso de dados industriais (e não de fluorescência). Neste caso, a base de dados utilizada para construção dos modelos preditivos parte da

simulação de três colunas de separação em série, inspiradas por uma unidade real de separação em operação. No estudo, a metodologia de seleção de variáveis AnTSbe é comparada com a aplicação direta de técnicas de regularização de modelos empíricos (Regressão *Ridge* e *LassoLars*). As três colunas de separação são tratadas individualmente, e as concentrações dos compostos de interesse são preditas tanto nas correntes de topo quanto de fundo, caracterizando o perfil de concentrações das colunas. Em especial, se foca na predição das impurezas de cada coluna, que devem ser mantidas ínfimas.

Por fim, o Capítulo 7 traz as conclusões e considerações gerais acerca do trabalho.

## 1.3 Produção Científica e Contribuições Paralelas

Os capítulos que compõem esta Tese de doutorado já foram ou serão submetidos à periódicos e congressos científicos. Ademais, durante a realização deste trabalho, o autor participou ativamente de projetos envolvendo análise de dados, que tem correlação direta com os conhecimentos que adquiriu durante a realização das pesquisas aqui expostas. Estão destacados abaixo os artigos completos publicados até a data desta defesa.

- C. Ranzan, L. Ranzan, L.F. Trierweiler, J.O. Trierweiler. *Sulfur Determination in Diesel using 2D Fluorescence Spectroscopy and Linear Models*. IFAC-PapersOnLine. 48 (2015) 415–420. (Congresso).

- L. Ranzan, C. Ranzan, L.F. Trierweiler, J.O. Trierweiler. *Classification of Diesel Fuel Using Two-Dimensional Fluorescence Spectroscopy*. Energy & Fuels. 31 (2017) 8942–8950.

- J.A. Sebben, J. da Silveira Espindola, L. Ranzan, N. Fernandes de Moura, L.F. Trierweiler, J.O. Trierweiler. *Development of a quantitative approach using Raman spectroscopy for carotenoids determination in processed sweet potato*. Food Chem. 245 (2018) 1224–1231.

- P.V.J.L. Santos, L. Ranzan, M. Farenzena, J.O. Trierweiler. *K-Rank: An Evolution of Y-Rank for Multiple Solutions Problem*. Brazilian J. Chem. Eng. 36 (2019) 409–419.

- D.G. Carvalho, L. Ranzan, L.F. Trierweiler, J.O. Trierweiler. Determination of the Concentration of Total Phenolic Compounds in Aged Cachaça Using Two-Dimensional Fluorescence and Mid-Infrared Spectroscopy. Food Chem. 329 (2020) 127142.

- Weber, C.T., Ranzan, L., Liesegang, L.L.M., L.F. Trierweiler, J.O. Trierweiler. *A circular economy model for ethanol and alcohol-based hand sanitizer from sweet potato waste in the context of COVID-19*. Brazilian Journal of Operations & Production Management, Vol. 17 (2020), No. 03, e20201025.

- D.G. Carvalho, L. Ranzan, R.A. Jacques, L.F. Trierweiler, J.O. Trierweiler. *Analysis of total phenolic compounds and caffeine in teas using variable selection approach with two-dimensional fluorescence and infrared spectroscopy*. Microchemical Journal 169(2) (2021),106570.

Na sequência, estão listados os Trabalhos de Conclusão de Curso em que o autor participou como coorientador, vários dos quais aplicam partes dos códigos e metodologias que o autor desenvolveu:

- Zanatta, Felipe Georg (2016). Potencial econômico e aplicação do bio-óleo da pirólise rápida de casca de arroz como pesticida. Orientadores: J. Trierweiler, L. F. Trierweiler. Coorientador: L. Ranzan. Trabalho de Diplomação em Engenharia Química, Departamento de Engenharia Química – UFRGS.

- Boff, Luiz Gustavo Fracalossi (2018). Estudo do potencial da utilização da espectroscopia para classificação de cervejas. Orientadores: J. Trierweiler, L. F. Trierweiler. Coorientador: L. Ranzan. Trabalho de Diplomação em Engenharia Química, Departamento de Engenharia Química – UFRGS.

- Trucolo, Ana Clara Fernandes (2018). Desenvolvimento de um analisador virtual para determinar o teor de umidade de protreínas isoladas de soja. Orientadores: J. Trierweiler, L. F. Trierweiler. Coorientador: L. Ranzan. Trabalho de Diplomação em Engenharia Química, Departamento de Engenharia Química – UFRGS.

- Freitas, Lucas do Amaral (2018). Estudo da viabilidade técnica na utilização de espectroscopia Raman na caracterização online de solos. Orientador: C. Ranzan. Coorientador: L. Ranzan. Trabalho De Diplomação Em Engenharia Agroindustrial - Agroquímica, Escola De Química e Alimentos – FURG.

- Gomes, Alexandre Rodrigues Conill (2019). Aplicação de análise multivariada em dados de agrotóxicos na água para consumo humano do Brasil de 2014 a 2018. Orientador: J. Trierweiler. Coorientadores: L. Ranzan, L. B. Zini. Trabalho de Diplomação em Engenharia Química, Departamento de Engenharia Química – UFRGS.

- Liesegang, Luciano Luís Menz (2020). Produção de álcool gel em biorrefinarias descentralizadas de etanol a partir de batata-doce. Orientadores: J. Trierweiler, L. F. Trierweiler. Coorientador: L. Ranzan. Trabalho de Diplomação em Engenharia Química, Departamento de Engenharia Química – UFRGS.

## 1.4 Contribuições

As principais contribuições desta tese podem ser listadas como:

- **$C_1$** AnTSbe: metodologia hibrida ACO – TS para seleção de variáveis e ajuste de modelos com expansão de bases. Implementada em software livre Python e que pode ser aplicada em diferentes bases de dados, e não só para fluorescência.

- **$C_2$** Adapted AnTSbe: Desenvolvimento da metodologia DeltaPair para aprimorar a generalização de modelos com expansão de base.

- **$C_3$** Mapeamento espectral e captura do perfil fenólico em diferentes processos de envelhecimento de cachaças.

- **$C_4$** Calibração de redes neurais convolucionais residuais com matrizes excitação – emissão de fluorescência coletados continuamente e aplicação bem-sucedida dos

modelos ajustados para predição de importantes constituintes em sistemas fermentativos.

- **$C_5$** Desenvolvimento da metodologia de triagem e identificação de dados anômalos baseado no erro de reconstrução de redes *autoencoder.*

- **$C_6$** Desenvolvimento, nos estudos de caso, de novas metodologias quantitativas baseadas em fluorescência e dados industriais, com potencial para ferramenta de projeto para sensores de processo.

## 1.5 Resumo Gráfico

A Figura 1.1 apresenta um resumo gráfico que correlaciona os objetivos com as contribuições específicas de cada um dos capítulos. Assim, é possível ter uma visão geral sobre a estrutura e interligações entre as diferentes etapas do trabalho.

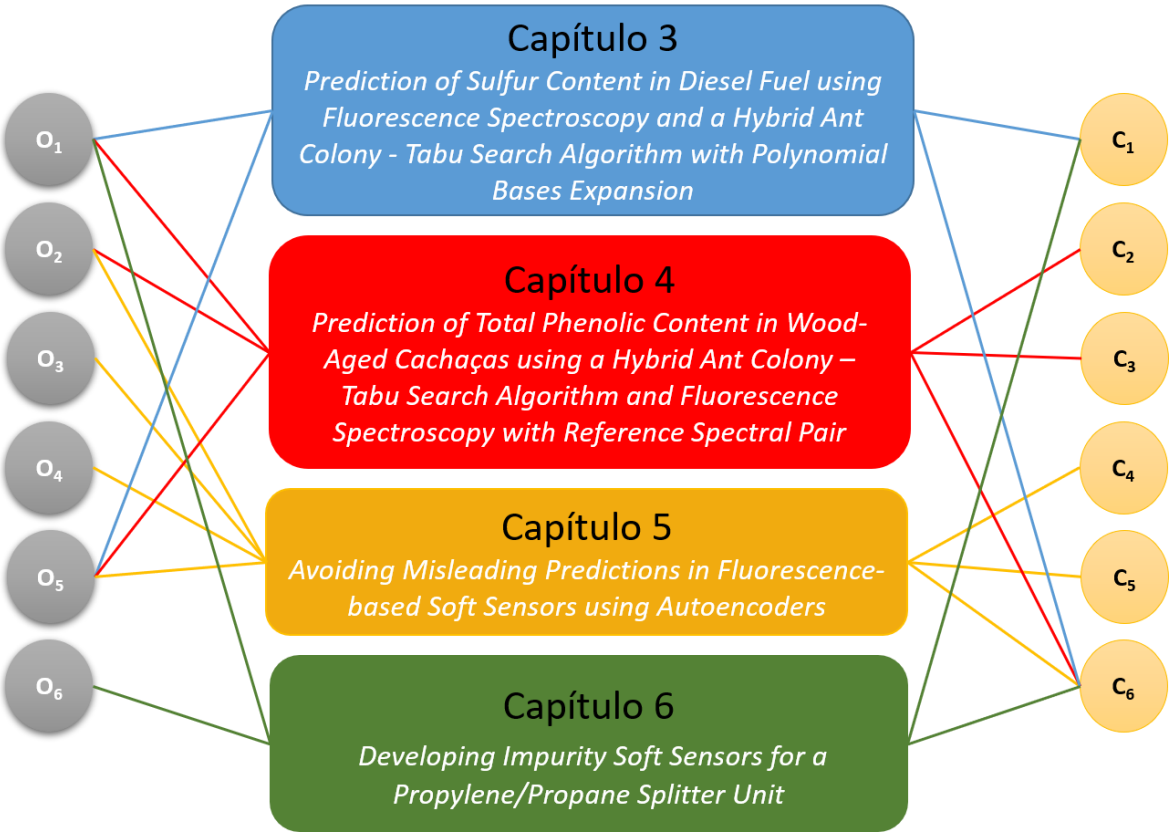



Figura 1.1. Resumo gráfico, indicando as correlações entre objetivos, capítulos e contribuições do trabalho.

## 1.6 Referências

ABIODUN, O. I. et al. **State-of-the-art in artificial neural network applications: A surveyHeliyon**Elsevier Ltd, , 1 nov. 2018.

ALMOTIRI, J.; ELLEITHY, K.; ELLEITHY, A. Comparison of Autoencoder and Principal Component Analysis Followed by Neural Network for E-Learning Using Handwritten Recognition. **2017 IEEE Long Island Systems, Applications and Technology Conference, LISAT 2017**, n. May, 2017.

AMARBAYASGALAN, T.; JARGALSAIKHAN, B.; RYU, K. Unsupervised Novelty Detection Using Deep Autoencoders with Density Based Clustering. **Applied Sciences**, v. 8, 2018.

BELTRAMO, T.; KLOCKE, M.; HITZMANN, B. Prediction of the biogas production using GA and ACO input features selection method for ANN model. **Information Processing in Agriculture**, v. 6, n. 3, p. 349–356, 2019.

BHATTACHARYYA, S. **Hybrid Metaheuristics**. [s.l.] WORLD SCIENTIFIC, 2018. v. 84

CHARTE, D. et al. A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. **Information Fusion**, v. 44, n. December 2017, p. 78–96, 2018.

DORIGO, M.; GAMBARDELLA, L. M. Ant colonies for the travelling salesman problem. **Biosystems**, v. 43, n. 2, p. 73–81, 1997.

EINIPOUR, A.; CORRESPONDING. A fuzzy-ACO method for detect breast cancer. **Global Journal of Health Science**, v. 3, 2011.

JES, F. et al. Deep Convolutional Autoencoders vs PCA in a Highly-Unbalanced Parkinson ' s Disease Dataset : A DaTSCAN Study. **Springer Nature**, p. 47–56, 2019.

LU, Y. **Industry 4.0: A survey on technologies, applications and open research issuesJournal of Industrial Information Integration**Elsevier B.V., , 1 jun. 2017.

MISRA, S. et al. **An Autoencoder Based Model for Detecting Fraudulent Credit Card Transaction**. Procedia Computer Science. **Anais**...Elsevier B.V., 1 jan. 2020

MORO, S.; RITA, P.; VALA, B. Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. **Journal of Business Research**, v. 69, n. 9, p. 3341–3351, 2016.

RANZAN, C. et al. **Validação com Dados de Espectroscopia Fluorescente 2D de Modelo Dinâmico para Fermentações Batelada de Saccharomyces cerevisiae**Caxias do Sul, RS, Brasil18 Simpósio Nacional de Bioprocessos, , 2011.

RANZAN, C. et al. Wheat flour characterization using NIR and spectral filter based on Ant Colony Optimization. **Chemometrics and Intelligent Laboratory Systems**, v. 132, n. 0, p. 133–140, 2014.

RANZAN, C. et al. Sulfur Determination in Diesel using 2D Fluorescence Spectroscopy and Linear Models. **IFAC-PapersOnLine**, v. 48, n. 8, p. 415–420, 2015.

RANZAN, L. et al. Classification of Diesel Fuel Using Two-Dimensional Fluorescence Spectroscopy. **Energy & Fuels**, v. 31, n. 9, p. 8942–8950, 2017.

RUIZ-DEL-SOLAR, J.; LONCOMILLA, P.; SOTO, N. A Survey on Deep Learning Methods for Robot Vision. 28 mar. 2018.

SAKURADA, M.; YAIRI, T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. **Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis**, p. 4, 2014.

SKOOG, D. A.; HOLLER, J. F.; CROUCH, S. R. **Princípios de Análise Instrumental**. 6ª edição ed. [s.l.] Bookman, 2009.

SUN, L. et al. A Hybrid Gene Selection Method Based on ReliefF and Ant Colony Optimization Algorithm for Tumor Classification. **Scientific Reports**, v. 9, n. 1, 2019.

TANG, J.; ALELYANI, S.; LIU, H. Feature selection for classification: A review. In: **Data Classification: Algorithms and Applications**. [s.l: s.n.]. p. 37–64.

ZOU, J.; ZHANG, J.; JIANG, P. Credit Card Fraud Detection Using Autoencoder Neural Network. 30 ago. 2019.

# Capítulo 2 – **Revisão Bibliográfica**

Neste capítulo consta o levantamento bibliográfico acerca dos principais assuntos constituintes desta tese. Por facilidade, o mesmo se encontra dividido em subseções. Primeiramente, serão abordadas as definições de modelos empíricos e suas peculiaridades, como o uso de normas de regularização. Na sequência, serão apresentados conceitos e metodologias aplicadas na seleção de variáveis e otimização de modelos. A seguir, uma visão geral sobre algoritmos do tipo redes neurais, com a definição de tipos específicos de redes, e os parâmetros das mesmas. Por fim, discorreremos sobre os fundamentos da espectroscopia por fluorescência, técnica na qual se baseiam os estudos de caso desta tese.

## 2.1 Modelos Empíricos

Modelos empíricos são aqueles baseados inteiramente em dados experimentais. Em Engenharia, são comumente denominados de modelos caixa-preta, descrevendo um sistema fechado de complexidade potencialmente alta, mas que se limita a medidas das relações entre estímulos de entrada e respostas de saída, sem que seja levada em consideração a estrutura interna (GUIDOTTI et al., 2018). Uma representação visual pode ser vista na Figura 2.1. Os modelos empíricos estabelecem uma relação funcional entre os sistemas de entrada e saída, onde os parâmetros do mesmo não necessitam ter significância fenomenológica (como coeficientes de transferência de massa e calor ou cinética de reação), mas conseguem, de maneira eficiente, representar as tendências de comportamento do processo (ZHANG, 2010).


Figura 2.1. Representação visual de um modelo caixa-preta.

As principais vantagens no uso de modelos empíricos se dão pela sua relativa simplicidade. Com uma base de dados suficientemente grande, os modelos podem correlacionar de maneira acurada entradas e saídas das quais os reais mecanismos

fenomenológicos não são conhecidos ou são muito complexos. Por se tratarem normalmente de relações matemáticas simples, os modelos podem ser facilmente ajustados a novos conjuntos de dados. Considerando a atual performance dos sistemas computacionais, são uma alternativa barata e rápida, tanto para uso quanto para implementação (HEIYANTHUDUWAGE; MOUNOURY; KOVACEVIC, 2011).

Porém, as vantagens apresentadas também carregam desvantagens importantes. Primeiramente, a base de dados experimental tem papel fundamental: a mesma precisa ser não somente vasta, como de qualidade. Isso quer dizer que muitos dados precisam ser coletados, com a maior variabilidade possível no processo, contemplando as variações de comportamento que se espera capturar com o modelo. Esta alta necessidade por dados pode ser proibitiva para sistemas de difícil coleta e/ou experimentação. O modelo ajustado geralmente tem baixo poder de generalização, perdendo potencial preditivo ao se afastar das condições operacionais similares as dos dados de treinamento. Assim, estratégias de retroalimentação e atualização dos modelos são importantes. Ademais, considerando que o modelo não é baseado em relações físicas, as equações e algoritmos finais nem sempre provêm compreensão sobre o processo (GUIDOTTI et al., 2018).

Supondo que nossos dados reais surjam de um modelo estatístico na forma da Equação 2.1:

$$Y = f(X) + \varepsilon \tag{2.1}$$

que relaciona o sistema observado entradas – saídas (X,Y) através da função $f$ e do erro randômico $\varepsilon$, independente de X (que captura erros de medição de X e variáveis não mensuradas que contribuem para Y). Assim, o objetivo da modelagem empírica é encontrar uma aproximação útil $\hat{f}(x)$ para função $f(x)$, capaz de representar a relação preditiva entre as observações de entrada e saídas. Para tal, o sistema em estudo é observado e um conjunto de calibração contendo tanto as entradas quanto as saídas na forma $\mathcal{T} = (x_i, y_i), i = 1, \dots, N$ é agrupado. Os valores de entrada também são alimentados em um sistema artificial (algoritmo de aprendizado), que produz uma saída $\hat{f}(x_i)$ em resposta as entradas. O algoritmo de aprendizado tem a capacidade de alterar a sua relação entrada/saída para responder as diferenças $y_i - \hat{f}(x_i)$, entre as saídas reais e as saídas preditas. Se espera que ao fim do processo de aprendizagem, as saídas reais e artificiais sejam suficientemente parecidas para representar de maneira útil todos os conjuntos de dados de entrada que possam ser encontrados na prática (HASTIE, TREVOR, TIBSHIRANI, ROBERT, FRIEDMAN, 2009).

São diversas as abordagens para definição da função de aproximação $\hat{f}(x)$, em sua maioria associando as variáveis de entrada a uma série de parâmetros $\theta$ que podem ser modificados para se adequar aos dados disponíveis . Entre as mais aplicadas, podemos citar o uso de funções polinomiais (OSTERTAGOVÁA, 2012), de regressões baseadas em transformações do espaço (como as regressões por mínimos quadrados parciais (PLS) (GELADI; KOWALSKI, 1986) e as regressões de componentes principais (PCR) (HEMMATEENEJAD; MIRI; ELYASI, 2012)), e uso de funções não lineares, incluindo modelos do tipo redes neurais (BASHEER; HAJMEER, 2000). Porém, as funções lineares ainda são umas das mais versáteis aproximações utilizadas (YU; YAO, 2017).

### *2.1.1 Regressões Lineares e Soma dos Quadrados dos Resíduos*

Considerando um vetor de entradas $\mathrm{X}^{\mathrm{T}} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_p)$, e com a intenção de predizer uma saída real $Y$, o modelo de regressão linear apresenta a forma da Equação 2.2.

$$f(X) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j \tag{2.2}$$

Onde $\beta_j$ são parâmetros desconhecidos e *p* é a dimensão do subespaço das variáveis de entrada. Desta forma, as variáveis $\mathrm{x}_j$ podem ser originarias de diversas fontes:

- Variáveis quantitativas;
- Transformações das variáveis quantitativas, como logaritmo, raiz quadrada ou potências;
- Representação polinomial das entradas, como $\mathrm{x}_2 = {\mathrm{x}_1}^2$ e $\mathrm{x}_3 = {\mathrm{x}_1}^3$;
- Interação entre as variáveis, como $\mathrm{x}_3 = \mathrm{x}_1 \times \mathrm{x}_2$.

Neste caso, independente da fonte de $\mathrm{x}_j$, o modelo é linear nos parâmetros.

A maneira mais tradicional de estimar os parâmetros $\beta$ é através do método dos Mínimos Quadrados, onde os coeficientes $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ são escolhidos a fim de minimizar a soma dos quadrados dos resíduos (RSS – *Residual Sum of Squares*)

$$RSS(\beta) = \sum_{i=1}^{N} \big(y_i - f(x_i)\big)^2$$

$$RSS(\beta) = \sum_{i=1}^{N} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \tag{2.3}$$

onde$(x_i, y_i)$ são as observações de entrada e saída do conjunto de calibração.

Para minimizar Eq. 2.3 podemos escrever o RSS como

$$RSS(\beta) = (y - X\beta)^T (y - X\beta) \tag{2.4}$$

Que é uma função quadrática nos *p* + 1 parâmetros. Derivando em relação a $\beta$, obtemos

$$\frac{\partial RSS}{\partial \beta} = -2X^T (y - X\beta) \tag{2.5}$$

$$\frac{\partial^2 RSS}{\partial\beta\partial\beta^T} = 2X^T\mathrm{X}$$

Assumindo que X tem posto de colunas completo, e, assim, $X^T\mathrm{X}$ é positivo definido, a primeira derivada pode ser zerada

$$X^T(y - X\beta) = 0 \tag{2.6}$$

e a solução única é obtida

$$\hat{\beta} = (X^T X)^{-1}X^T y \tag{2.7}$$

O modelo de predição ajustado é assim representado por

$$\hat{f}(x_i) = X\hat{\beta} \;=\; X(X^T X)^{-1}X^T y \tag{2.8}$$

Por se tratar de uma resposta fechada, única e de fácil obtenção (dependendo basicamente de inversão e transposição de matrizes), a abordagem de modelos lineares e ajuste por mínimos quadrados é conveniente desde antes do uso de computadores.

### *2.1.2 Normas de Regularização*

Quando variáveis correlacionadas estão presentes em modelos lineares, os coeficientes podem acabar mal determinados e apresentar alta variância. Um coeficiente positivo grande em uma variável pode ser cancelado por um coeficiente negativo similar em uma variável correlacionada. Assim, técnicas de penalização foram desenvolvidas para encolher ou remover os coeficientes dos modelos.

A Regressão Ridge, também conhecida como norma $l_2$, reduz os coeficientes da regressão impondo uma penalização quadrática referente ao seu tamanho,

$$\hat{\beta}^{ridge} \;=\; \underset{\beta}{argmin}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}{\beta_j}^2\right\} \tag{2.9}$$

onde $\lambda \geq 0$ é um parâmetro referente a força da penalização. Com alguma manipulação matemática, a solução para a regressão Ridge pode ser escrita como

$$\hat{\beta}^{ridge} = (X^T X + \lambda\mathrm{I})^{-1}X^T y \tag{2.10}$$

onde I é a matriz identidade $p\ x\ p$. A seleção da penalização quadrática $\beta^T\beta$ faz com que a solução da regressão Ridge seja novamente uma função linear de $y$. Esta solução adiciona uma constante positiva a diagonal $X^T X$ antes da inversão, fazendo com que o problema se torne não singular, mesmo quando a matriz não possui posto completo. Esta foi a principal característica estudada pelos desenvolvedores da metodologia (HOERL;

KENNARD, 1970). Por sua característica quadrática, a norma $l_2$ faz com que os coeficientes tendam a zero, mas não podem ser explicitamente zero.

A metodologia Lasso, conhecida como norma $l_1$, é definida em Eq 2.11

$$\hat{\beta}^{lasso} = \underset{\beta}{argmin} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left( y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.11}$$

Neste caso, a penalidade $l_2$ Ridge $\sum_1^p {\beta_j}^2$ é substituída pela penalidade $l_1$ Lasso $\sum_1^p |\beta_j|$. Esta restrição torna as soluções não lineares em $y_i$, e, assim, não existe uma expressão de forma fechada como na regressão Ridge. A solução para o Lasso se torna um problema de computação quadrática, porém, existem algoritmos disponíveis que tornam a computação de todo o caminho de soluções com a variação de $\lambda$ de custo computacional comparável aquele da regressão Ridge, como por exemplo a metodologia LAR (*Least Angle Regression* (EFRON et al., 2004)). Dada a natureza da penalização $l_1$ os coeficientes das variáveis podem ser reduzidos a exatamente zero.

A Figura 2.2 apresenta uma comparação entre as regularizações Ridge (a direita) e Lasso (a esquerda). A RSS, em vermelho, apresenta contornos elípticos, centrados na estimativa de mínimos quadrados. A áreas em azul representam as regiões das restrições. A solução de ambos os métodos se dá no primeiro ponto de contato entre as regiões e as elipses.

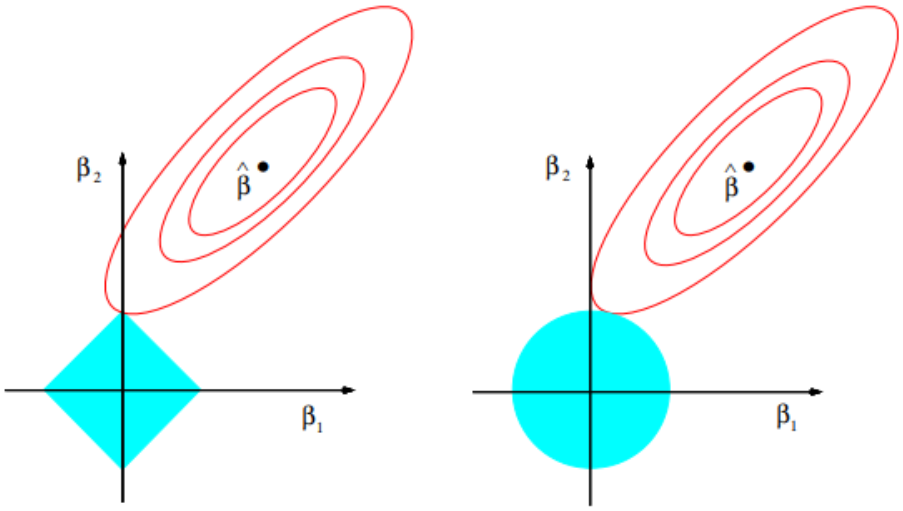



Figura 2.2. Comparação entre as penalizações Lasso (esquerda) e regressão Ridge (direita). A área em azul representa as regiões de restrição $|\beta_1| + |\beta_2| \leq t$ e $\beta_1^2 + \beta_2^2 \leq t^2$, respectivamente. As elipses vermelhas representam o contorno da função erro dos mínimos quadrados. Fonte: (HASTIE, TREVOR, TIBSHIRANI, ROBERT, FRIEDMAN, 2009)

## 2.2 Otimização de Modelos e Seleção de Variáveis

Em muitas aplicações, o número de variáveis de entrada disponíveis em um sistema pode atingir a casa das centenas até os milhares (por exemplo, classificação de imagens (JOHNSON; XIE, 2013; KAVZOGLU, 2017; MOHAPATRA; PATRA; SATPATHY, 2014), dados espectrais (RANZAN et al., 2014, 2017; SEBBEN et al., 2018) e processos industriais (SILVA; SECCHI, 2018; ZHENG et al., 2012)). Todavia, esta abundância de dados muitas vezes não apenas não apresenta vantagens para predizer o sistema, como ativamente causam um detrimento dos modelos preditivos. Estes dados podem estar associados a um alto nível de ruído, de colinearidade, e repletos de variáveis redundantes e/ou irrelevantes para a propriedade que se pretende inferir (TANG; ALELYANI; LIU, 2014).

Diversos são os estudos focados no desenvolvimento de técnicas capazes de extrair a informação válida das variáveis de entrada de um sistema, reduzindo o ruído e as entradas sem serventia, mas mantendo todas as relações e padrões importantes do conjunto de dados (BALABIN; SMIRNOV, 2011; BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2013; CHANDRASHEKAR; SAHIN, 2014; PES, 2019; XIAOBO et al., 2010). Este processo de seleção de variáveis é considerado essencial para a otimização de modelos empíricos, sendo muitas vezes a diferença entre modelos falhos de modelos efetivos. Amplamente, estas técnicas podem ser categorizadas em Extração de Características (*Feature Extraction*) e Seleção de Variáveis (*Feature Selection*).

A abordagem de Extração de Características visa projetar as variáveis em um novo espaço com menor dimensionalidade, através da combinação do espaço original das mesmas. O novo espaço tem a intenção de ser informativo, não redundante e facilitar o aprendizado e generalizações a partir dos dados. Muitas vezes, a análise deste novo espaço facilita a interpretação e a busca por padrões nos dados. Os principais exemplos destas técnicas incluem Análise de Componentes Principais (*Principal Component Analysis* – PCA) (CAMACHO; PICÓ; FERRER, 2010; JACKSON, 1991; JOLLIFFE, 1986; TIPPING; BISHOP, 1999), Análise Discriminante Linear (*Linear Discriminant Analysis* – LDA) (SILVA et al., 2016), Regressão por Mínimos Quadrados Parciais (*Partial Least Squares* – PLS) (CHI et al., 2014; FILZMOSER; TODOROV, 2011; GELADI; KOWALSKI, 1986; WOLD; SJÖSTRÖM; ERIKSSON, 2001; XU et al., 2007), e Análise de Fatores Paralelos (PARAFAC) (MURPHY et al., 2013; SILVA et al., 2019).

As técnicas de PCA e regressão PLS são amplamente utilizadas para análise de dados (GODOY; VEGA; MARCHETTI, 2014). Estas técnicas são de fácil aplicação, estando implementas em praticamente todos os softwares matemáticos modernos. Seu custo computacional também é irrisório, sendo sua aplicação praticamente instantânea. Na PCA, a matriz original dos dados **X** é decomposta como produto de duas outras matrizes, **T** (denominada *scores*) e **P** (denominada *loadings*), somada de uma matriz de residuais **E**

$$\mathrm{X} = \mathrm{TP}' + E = \sum_{a=1}^{A} t_a \, p_a' + E \tag{2.12}$$

O número de colunas da matriz **T** e linhas da matriz **P** correspondem a quantidade componentes principais (PCs), e onde cada $t_a$ representa um PC. Os PCs são ortogonais entre si, e equivalem ao novo espaço formado a partir da combinação linear do espaço original. De acordo com as características da decomposição, a direção ortogonal de cada

PCs é definida a fim de capturar a máxima variância contida nos dados. Assim, a quantidade de informação independente contida nos dados é acumulada em ordem decrescente em cada PC, até o ponto em que a adição de mais PCs na matriz não mais adiciona informação (variância). Desta forma é realizada a redução do espaço, truncando o número de PCs naquele capaz de capturar suficientemente a variância original dos dados.

A PCA é uma análise não supervisionada, ou seja, considera apenas os dados de entrada sem nenhuma relação com as saídas do processo. Já a regressão PLS consiste de duas correlações externas que unem os resultados das decomposições simultâneas tanto da matriz que contém as variáveis de entrada **X** quanto a matriz contendo as saídas **Y**, e uma correlação interna entre as matrizes de scores de ambas decomposições. O objetivo da modelagem PLS é minimizar a norma de **F** ao mesmo tempo que maximiza a covariância entre **X** e **Y**. Esta relação interna é uma regressão linear múltipla entre as matrizes de scores **U** e **T** em que **B** é uma matriz de coeficiente de regressão $n \times n$, determinada por minimização de mínimos quadrados (ROBERTSON et al., 2009). Neste caso, cada nova dimensão do espaço projetado é denominada de variável latente, e elas não necessariamente são ortogonais entre si. Uma simplificação visual da forma matricial da regressão PLS pode ser vista na Figura 2.3. Podemos perceber pela figura que, se realizadas individualmente, as decomposições das matrizes **X** e **Y** equivalem a uma PCA.

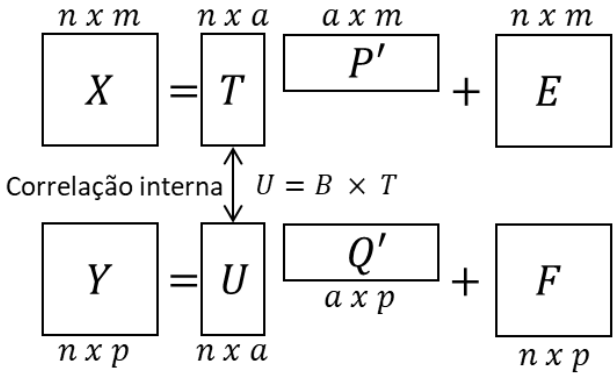



Figura 2.3. Simplificação visual da forma matricial da regressão PLS.

Apesar da grande funcionalidade, as técnicas de Extração apresentam também algumas desvantagens. Primeiramente, o novo espaço criado, apesar de reduzido, ainda é formado pela combinação de todo o espaço original. Sendo assim, se mantem a necessidade de capturar toda a informação original. Se o objetivo da redução de dimensão for reduzir também a necessidade de coleta de dados, isto não acontece com a aplicação direta destas técnicas. Além disso, a combinação faz com que a significância física das variáveis originais seja perdida.

Por outro lado, as abordagens de Seleção de Variáveis visam a seleção de um pequeno subconjunto dos dados originais capaz de reduzir a redundância e maximizar a relevância em relação a função objetivo. Desta forma, as significâncias físicas das variáveis são mantidas e a redução de dimensão acarreta diretamente na redução da necessidade de coleta de dados (TANG; ALELYANI; LIU, 2014). Os métodos de Seleção podem ser categorizados em três grupos (PES, 2019):

(I) *Filters.* São técnicas que se baseiam nas características dos dados. O processo de seleção é conduzido em uma etapa de pré-processamento, sem interagir diretamente com o algoritmo de aprendizado usado na etapa de construção dos modelos (*e.g.,* Fisher Score (O. DUDA; E. HART; G.STORK, 2019), ReliefF (ROBNIK-ŠIKONJA; KONONENKO, 2003)).

(II) *Wrapper.* Usam o preditor como um modelo caixa-preta e a performance do preditor como a função objetivo para avaliar o subconjunto de variáveis selecionadas. O componente de seleção de variáveis escolhe um subconjunto que é fornecido ao algoritmo de aprendizado para predizer a saída de interesse. A performance do preditor é então devolvida ao componente de seleção de variáveis para ser utilizado na próxima iteração de seleção (TANG; ALELYANI; LIU, 2014). Se trata de uma metodologia iterativa, podendo ser conduzida até o limite onde todas as combinações possíveis de subgrupos sejam avaliadas, se tornando uma Busca Exaustiva. Para muitos casos, porém, o teste de todas as possíveis combinações é infactível. Assim, algoritmos de seleção sequencial (como *Sequential Feature Selection* (SFS) e seus variantes (REUNANEN; GUYON; ELISSEEFF, 2003)) e algoritmos de meta-heurísticas (como Algoritmo Genético (GOLDBERG, 1989), Otimização Enxame de Partículas (HU et al., 2019), Otimização Colônia de Formigas (PESSOA et al., 2015), Recozimento Simulado (KIRKPATRICK; GELATT; VECCHI, 1983) e Busca Tabu (MELO, 2008)), que produzem mínimos locais (e não globais), são utilizados.

(III) *Embedded.* O processo de seleção se baseia na capacidade intrínseca do algoritmo de aprendizado em designar pesos para as variáveis, sem utilizar um mecanismo de busca sistemática entre diversos subconjuntos. O objetivo destes métodos é reduzir o tempo computacional requerido pelos métodos *Wrapper*, aceitando uma troca entre a performance preditiva do modelo final, embutindo a etapa de seleção de variáveis dentro do próprio algoritmo de aprendizado (PES, 2019). Podemos citar como exemplo SVM-RFE (GUYON et al., 2002) e técnicas de regularização baseadas em norma $L_1$ como regularização Lasso e regularização Rede Elástica (TANG; ALELYANI; LIU, 2014).

### *2.2.1 Meta-heurísticas de Otimização*

As meta-heurísticas combinam conhecimento histórico dos resultados anteriores com escolhas aleatórias para guiar o processo de seleção de variáveis. Geralmente são aplicados para resolver problemas de otimização combinatória para os quais não se conhece um algoritmo específico eficiente. Cada meta-heurística aborda de maneira diferente os mecanismos de prospecção e exploração da superfície de resposta e a forma de qualificar e combinar os subconjuntos a cada iteração.

Nos Algoritmos Genéticos (GA), a população de possíveis soluções é tratada como vetores (cromossomos) formados por genes, e atua pela simulação do que seria a reprodução sexual. A cada iteração, os indivíduos resultantes são formados pela combinação das soluções contidas em seus pais. A qualidade dos descendentes é avaliada (aptidão), e indivíduos com baixa aptidão tem menor probabilidade de reprodução. Desta forma, as gerações subsequentes serão formadas por soluções cada vez melhores (GOLDBERG, 1989).

Os algoritmos de Enxame de Partículas (PSO) são inspirados pelo comportamento de bando (pássaros, peixes e insetos) e foram inicialmente propostos para resolução de problemas contínuos. As soluções em potencial (partículas) são iniciadas aleatoriamente e a melhor partícula é selecionada (qualidade da função objetivo). A cada iteração, as partículas voam através do espaço do problema seguindo na direção da partícula com a melhor solução atual. A direção do movimento a cada nova iteração depende tanto da posição da melhor partícula no enxame, quanto da posição da própria partícula. Regras estocásticas são introduzidas para dar um caráter aleatório ao algoritmo. O PSO se baseia na trajetória das partículas e nos pontos do espaço de busca visitados para qualificar a qualidade da solução, preservando os melhores locais visitados em uma estrutura de memória (KENNEDY; KENNEDY; EBERHART, 1995).

A Busca Tabu (TS), por sua vez, aplica métodos de busca local para otimizar modelos. Uma solução é iniciada (aleatória ou direcionada por algum método) e seus vizinhos imediatos são avaliados (soluções similares a atual, com a troca de pequenos detalhes). Se algum dos vizinhos possuir uma solução melhor que a atual, ela se torna a nova melhor solução do problema, e a solução anterior é adicionada à lista tabu, se tornando um movimento proibido. Para evitar ficar presa rapidamente em mínimos locais, a TS aplica critérios que permitem que a solução se mova em direções de piores respostas que a atual, desde que nenhum movimento melhor esteja disponível. Como a resposta anterior é incluída na lista tabu, o algoritmo evita tomar movimentos repetitivos e retornar à solução anterior, ampliando a região de busca. Regras também podem ser introduzidas para que movimentos (como combinação de certas variáveis) sejam considerados proibidos e assim não considerados durante a escolha dos vizinhos. Os algoritmos de TS geralmente trabalham com três tipos de memórias tabu: de curta, média e longa duração. A memória de curta duração diz respeito a movimentos realizados nas ultimas *t* iterações, e os mesmos se tornam novamente permitidos com a evolução da otimização. A memória de média duração pode ser utilizada para retornar à otimização a um movimento passado de boa solução, e criar um novo nó na busca, proibindo o caminho anteriormente escolhido. A memória de longa duração pode ser utilizada para recomeçar o algoritmo em um outro ponto da região de busca, aumentando a diversificação. Uma vez atingido o critério de parada do algoritmo, a melhor solução dentre todos os movimentos é apresentada como solução final (GLOVER, 1989, 1990).

A Otimização Colônia de Formigas (*Ant Colony Optimization* – ACO) é um algoritmo multiagente usado para resolver problemas de otimização combinatória NP-complexos (BRUCKER, 1979). Os algoritmos ACO tentam imitar o comportamento de forrageamento de formigas, transformando seu sistema de comunicação em um modelo estatístico. Em uma colônia real, a experiência coletiva das formigas durante a busca por alimento é transmitida através de traçadores de feromônio. A Figura 2.4 apresenta um exemplo desde comportamento, proposto pelos autores da metodologia original (DORIGO; GAMBARDELLA, 1997).

Nos algoritmos ACO, esse sistema é transformado em um vetor memória, e é representado por uma trilha de feromônios. Esta trilha governa a maneira com que os agentes navegam pelo espaço de busca e resumem suas experiências. No caso da seleção de variáveis e otimização de modelos, cada variável de entrada inicia a rotina de otimização

com uma mesma quantidade de feromônio. A cada iteração, cada uma das formigas seleciona um subconjunto das variáveis de entrada. Essa seleção se baseia em um componente randômico e um componente relacionado a quantidade de feromônio nas variáveis. A formiga entrega para o algoritmo de aprendizado o subconjunto, que ajusta um modelo para predizer as saídas de interesse. A performance do modelo é quantificada, e cada formiga deposita nas variáveis que selecionou uma quantidade de feromônio condizente: quanto melhor a performance, maior a quantidade de feromônio. Por fim, toda a trilha é evaporada, multiplicando a mesma por um valor entre 0 – 1, para penalizar variáveis não selecionadas. Esta operação garante um aumento da probabilidade, nas sucessivas iterações, da seleção de variáveis de entrada que participaram de modelos de alta qualidade, otimizando o preditor.

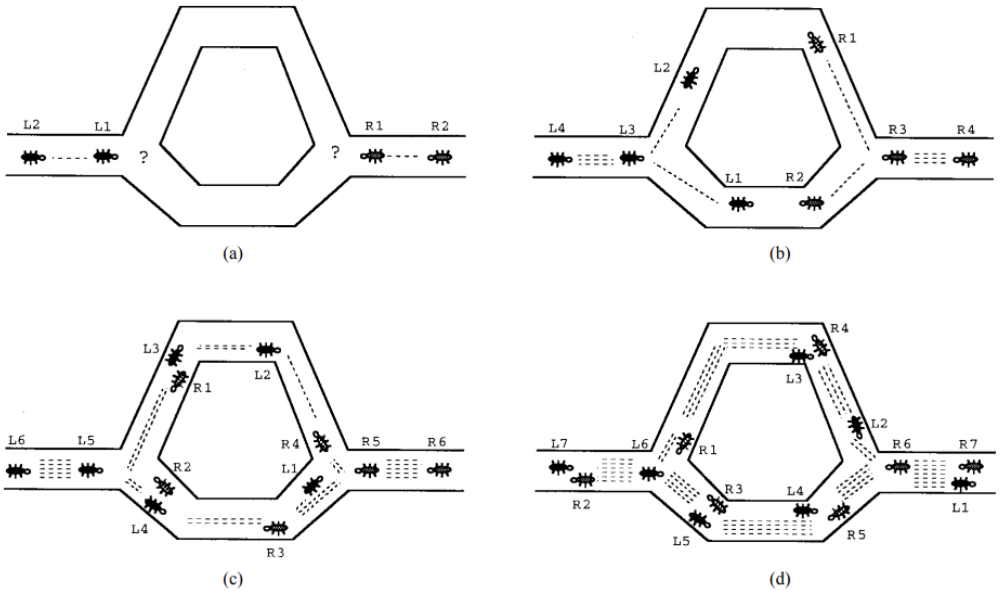



Figura 2.4. Definição de como formigas reais escolhem o menor caminho. (a) As formigas chegam ao ponto de decisão. (b) Randomicamente decidem o caminho a seguir. (c) As formigas que escolheram o caminho mais curto chegam ao ponto oposto em menos tempo. (d) O feromônio acumula no caminho mais curto em uma taxa mais alta. Assim, mais formigas tendem a seguir por ele. As linhas pontilhadas representam uma aproximação proporcional do feromônio depositado pelas formigas. Fonte: (DORIGO; GAMBARDELLA, 1997).

Na literatura, podemos encontrar diversos artigos de revisão que acompanham o desenvolvimento de metodologias baseadas em ACO. Reunindo os trabalhos de Dorigo e Blum (2005), Stützle, López-Ibáñez e Dorigo (2011) e Mohan e Baskaran (2012), são apresentadas mais de uma centena de pesquisas que aplicam ou modificam a metodologia ACO, em diversas áreas de estudo, como: topologia e otimização estrutural em eletrônica; aerodinâmica e dinâmica de fluidos; telecomunicações, bioinformática; finanças; modelagem, simulação e identificação de sistema em química, física e biologia; controle e processamento de sinais e imagens; problemas de roteamento, programação e problemas de produção, logística, transporte e gestão da cadeia de abastecimento.

A metodologia continua ativamente relevante, uma vez que permite muita liberdade para adaptações. Como uma evolução natural, é comum encontrar propostas de hibridização entre ACO e outras meta-heurísticas ou técnicas de otimização, para compensar por desvantagens na metodologia original (como, por exemplo, problemas de

*overfitting*, alta demanda computacional, estagnação precoce e ciclagem de subconjuntos). Algoritmos híbridos ACO já foram propostos para resolver Problemas de Caixeiro Viajante (ZHANG; TANG, 2008), Problema de Roteamento de Veículos (ABDULKADER; GAJPAL; ELMEKKAWY, 2015; BALSEIRO; LOISEAU; RAMONET, 2011; LI; TIAN; LEUNG, 2009; WANG et al., 2020), Problema De Atribuição De Tarefas Para Estações De Trabalho (SERBENCU; MINZU, 2016), Problema De Escalonamento De Projetos Com Restrição De Recursos (MYSZKOWSKI et al., 2015), entre outros. Um aprofundamento e compilação de estudos sobre o uso de meta-heurísticas híbridas pode ser visto nos trabalhos de revisão apresentados por Blum *et al.* (2008), Blum *et al.* (2011) e Bhattacharyya (2018).

## 2.3 Redes Neurais Artificiais

Redes neurais artificiais, como o nome sugere, são modelos matemáticos inspirados na estrutura básica de um cérebro. Este modelo consiste de diversas unidades de processamento (neurônios), operando paralelamente em uma camada, que são conectados aos neurônios de outras camadas por ligações (sinapses), representadas por pesos. Algumas vezes chamados de aproximadores universais, teorizasse que com unidades e informações suficientes os modelos possam aproximar virtualmente qualquer função a qualquer grau de precisão (CYBENKO, 1989; HORNIK; STINCHCOMBE; WHITE, 1989). Uma representação da estrutura de um neurônio artificial pode ser vista na Figura 2.5.

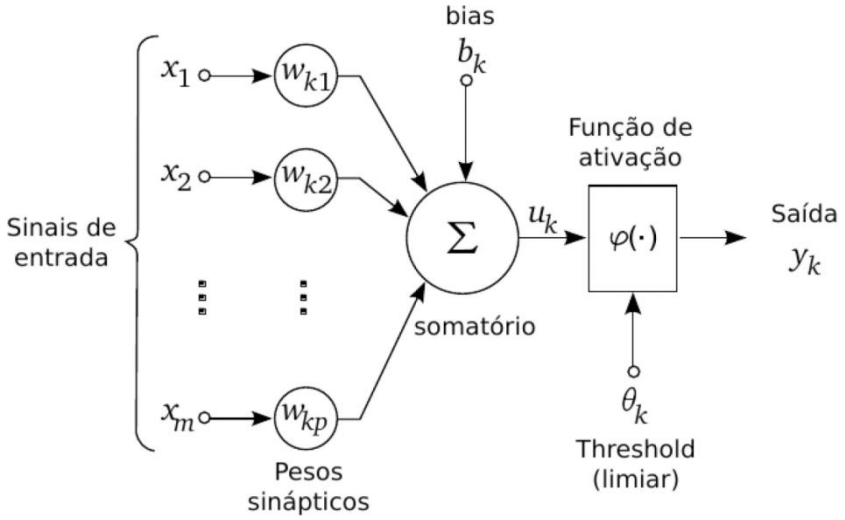



Figura 2.5. Estrutura básica de um neurônio artificial. Fonte: Schimidt *et al.* (2016).

A estrutura de uma rede neural diz respeito à disposição dos neurônios em camadas, e em como estas camadas estão conectadas entre si. As camadas são segmentadas em entradas, ocultas e saídas (sendo as camadas ocultas um número qualquer de camadas entre as entradas e saídas). Uma rede com apenas uma camada oculta é denominada rasa, e o acréscimo de mais camadas ocultas dá origem as chamadas redes profundas (*deep neural network*). O conceito de redes profundas também evoluiu junto com o desenvolvimento de metodologias mais eficientes de treinando das redes e do poderio computacional, sendo hoje possível treinar redes com centenas de camadas ocultas de maneira eficiente em poucos minutos (LEIJNEN; VAN VEEN, 2020).

Os neurônios, por sua vez, podem estar complemente interconectados com todos os neurônios da camada anterior e os da camada seguinte, formando a chamada unidade densa, ou o seu antônimo, chamada de unidade esparsa, onde uma seleção interna da rede poda conexões entre neurônios que não agregam informação útil (pesos próximos ou exatamente zero) (SRINIVAS; SUBRAMANYA; BABU, 2016).

Ainda sobre a conexão entre camadas da rede, a arquitetura mais comum é aquela onde a saída de uma camada é a entrada da camada subsequente. Este tipo de rede se denomina *feedforward*, e a informação é passada adiante de forma sequencial, sem que a função de saída da unidade tenha dependência com a própria saída de nenhuma forma. Em contraponto, redes do tipo recorrentes/*feedback* apresentam conexões em laços ou *loops*, onde a saída de camadas avançadas retorna às camadas anteriores (SCHWENK; BENGIO, 2000). A Figura 2.6 apresenta diferentes arquiteturas de redes neurais, e sua denominação usual.

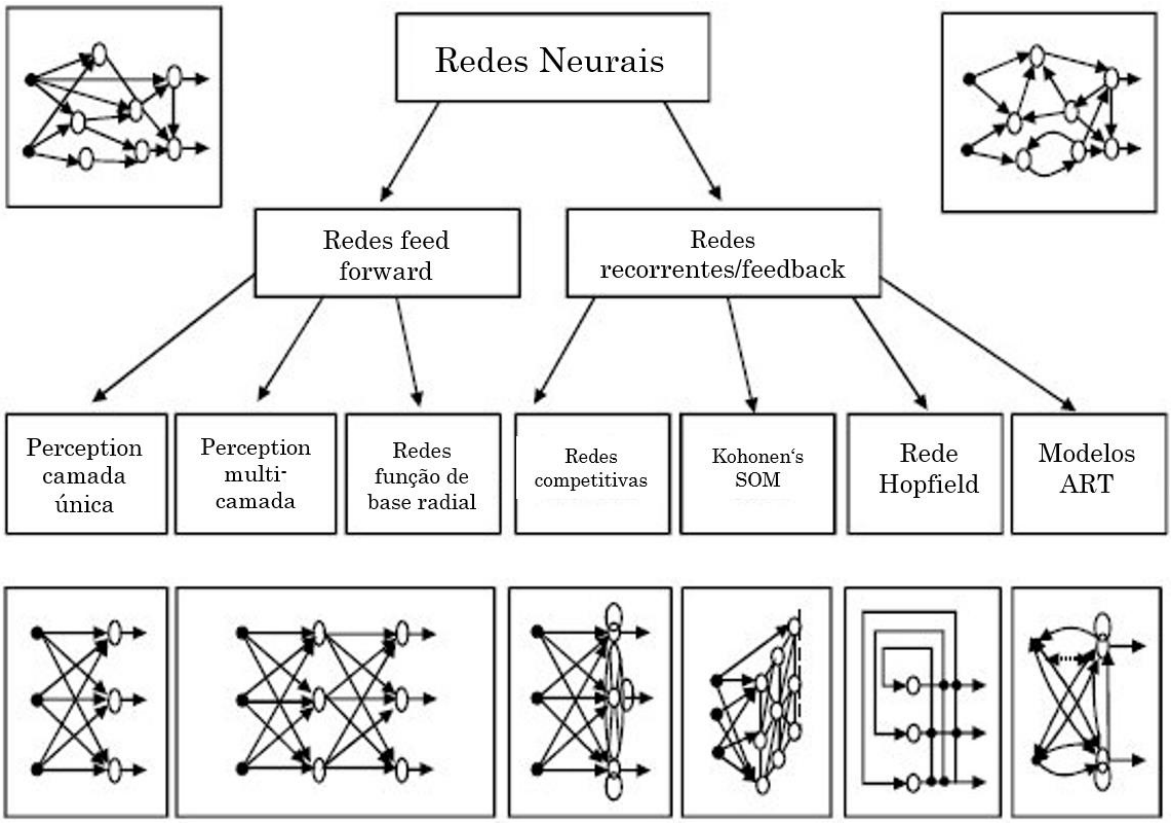



Figura 2.6. Diferentes arquiteturas de redes neurais artificias. Adaptada de (VERMA; KUMAR SINGH, 2015).

O principal conceito utilizado para o aprendizado das redes é o algoritmo de *backpropagation*, buscando a minimização de uma função objetivo para o ajuste de determinadas saídas. O algoritmo relaciona alterações nos parâmetros de peso e bias da rede às alterações na função objetivo. Essas alterações são calculadas iniciando pela última camada e seguem recursivamente até a primeira. As direções das alterações podem ser calculadas através das derivadas parciais dos erros, buscando a melhor alteração nos parâmetros capaz de minimizar a função objetivo. A magnitude da alteração nos parâmetros a cada iteração é comumente denominada taxa de aprendizado. Durante o processo de aprendizado pode ocorrer a saturação do neurônio, a depender do tipo de função de ativação e das entradas do mesmo. Neste caso, sua saída acaba sendo muito próxima de 0 ou 1, e alterações no peso pouco influenciem os outros neurônios da rede, e

por fim, a função objetivo. A saturação de neurônios (estado em que um neurônio emite valores predominantemente próximos às extremidades assintóticas da função de ativação) causa lentidão no aprendizado, principalmente quando o algoritmo de gradiente descendente é utilizado para otimização dos parâmetros (WIDROW; LEHR, 1990).

Um neurônio representa o somatório das unidades anteriores da rede, acrescido de um bias, sobre o qual é aplicada uma operação matemática, denominada função de ativação, gerando o sinal de saída da unidade. A função de ativação, assim, realiza transformações lineares ou não lineares nos sinais de entrada, permitindo que as mudanças nos pesos da rede sejam reguladas e que a rede consiga aprender a resolução para problemas complexos (CHARTE et al., 2018). Na literatura são sugeridas dezenas de abordagens para funções de ativação, sendo a função sigmoidal a primeira de grande destaque e as mais modernas variantes de unidades lineares como: ReLU, PReLU, ELU e SELU (ZHANG; LI, 2018). A Figura 2.7 apresenta algumas das mais comuns funções de ativação.

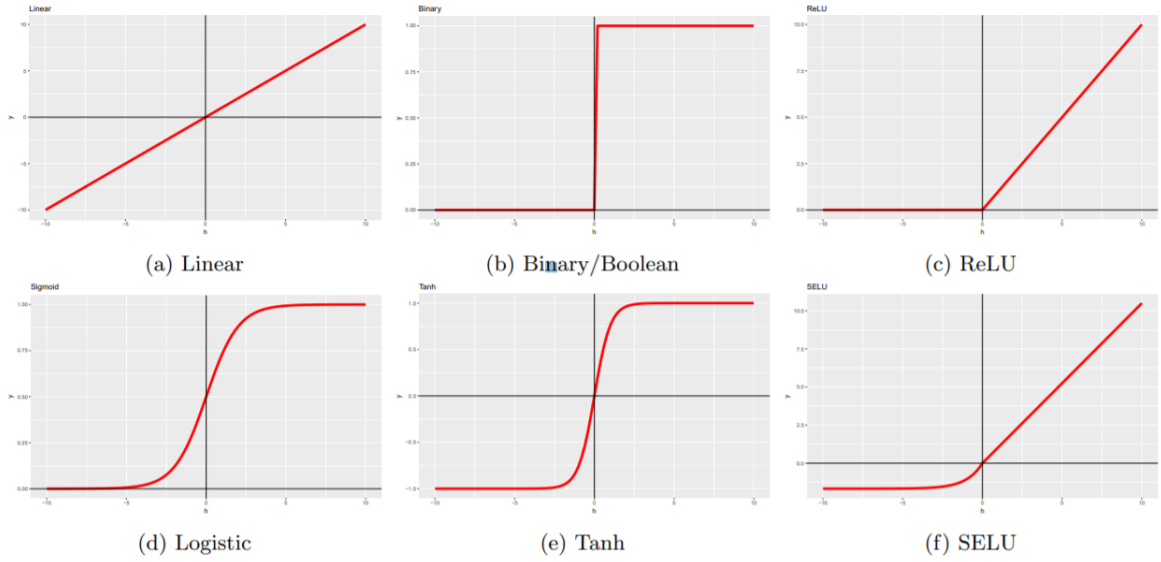



Figura 2.7. Funções de ativação comuns em redes neurais artificiais. Fonte: (CHARTE et al., 2018).

De maneira geral, as centenas de arquiteturas de redes propostas na literatura são combinações entre número de camadas, número de neurônios, e a maneira com que estes neurônios estão conectados entre si. Muitas destas arquiteturas são idealizadas com o objetivo de lidar com uma situação específica, como as redes convolucionais (CNN) e redes autoencoders (AE). As CNN agregam informações espaciais a rede, especialmente úteis para lidar com dados na forma de tensores (múltiplas dimensões, como imagens) (ZEILER; FERGUS, 2014). As redes AE, por sua vez, tem por objetivo treinar uma rede capaz de extrair de maneira não supervisionada padrões dos dados, sendo a rede capaz de codificar o espaço original em um novo espaço, e decodificar o mesmo buscando a reconstrução dos dados originais (CHARTE et al., 2018).

### 2.3.1 *Redes Neurais Convolucionais*

Redes convolucionais são consideradas por alguns pesquisadores como os melhores algoritmos de aprendizado para tratamento de imagens (CIREŞAN et al., 2012; LIU; DENG; YANG, 2019). Sua principal característica é sua habilidade em explorar correlações temporais e espaciais nos dados. A topologia da rede é dividida em múltiplos estágios compostos da combinação entre camadas convolucionais, unidades de processamento não linear e camadas de subamostragem (*pooling*) (LECUN; KAVUKCUOGLU; FARABET, 2010). As operações convolucionais extraem recursos uteis de dados localmente correlacionados. A saída da camada convolucional é transformada por uma função de ativação não linear, facilitando o aprendizado de diferenças semânticas nos dados. A saída da função de ativação é geralmente seguida de uma operação de subamostragem, que condensa os resultados e torna as entradas menos propensas a distorções geométricas (SCHERER; MÜLLER; BEHNKE, [s.d.]). Após os múltiplos estágios a rede é geralmente conectada a camadas densas de neurônios, para então gerar a saída de interesse. A Figura 2.8 apresenta um exemplo de CNN aplicada na classificação de imagens.

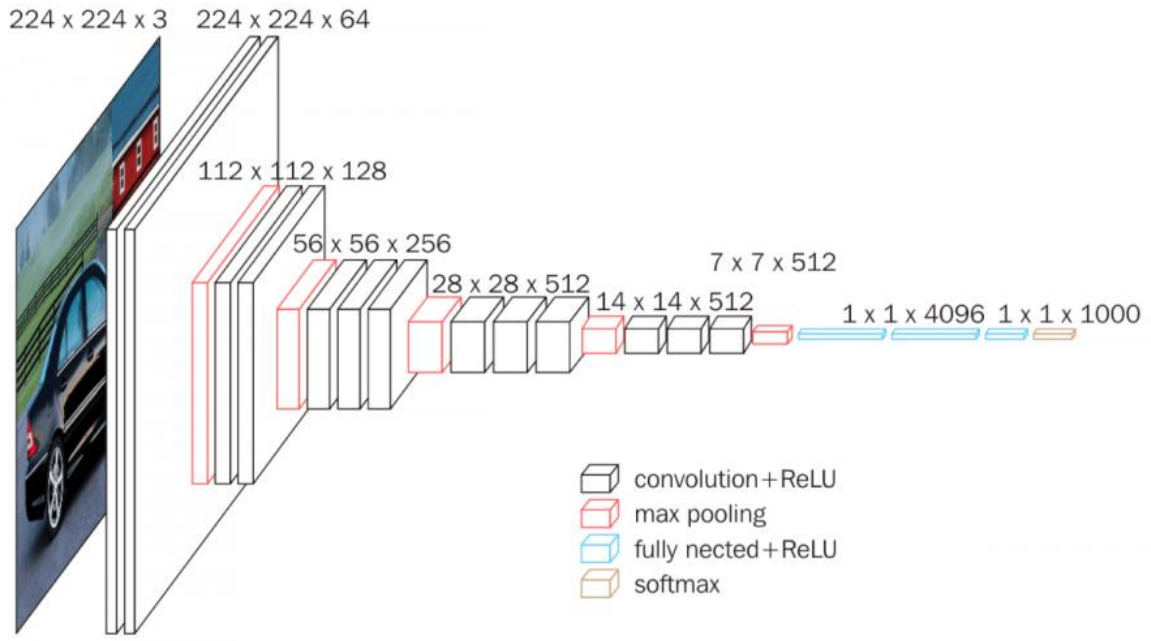



Figura 2.8. Exemplo de CNN aplicada na classificação de imagens. A imagem original possui 224x224 pixels e três canais (*Red-Blue-Green*). O espaço original é reduzido por camadas *max pooling*. A terceira dimensão apresentada é referente ao número de *kernels* (filtros) utilizados na camada convolucional. Fonte: (SIMONYAN; ZISSERMAN, 2015).

A arquitetura de CNN inclui diversos blocos construtivos, explicados a seguir.

#### 2.3.1.1 *Camadas Convolucionais*

Componente fundamental da arquitetura CNN, especializada na extração de características. Consiste de uma operação linear onde um pequeno número de matrizes, denominadas kernel, são aplicados no tensor de entrada. O produto elemento-a-elemento entre cada elemento do kernel e o tensor é calculado, em cada localização do tensor, e somado para obter o valor de saída naquela posição do tensor de saída, denominado mapa de recursos. O procedimento é aplicado repetidamente utilizando múltiplos kernels, formando um número arbitrário de mapas de recursos, que capturam diferentes características do tensor. A Figura 2.9 representa a operação convolucional.

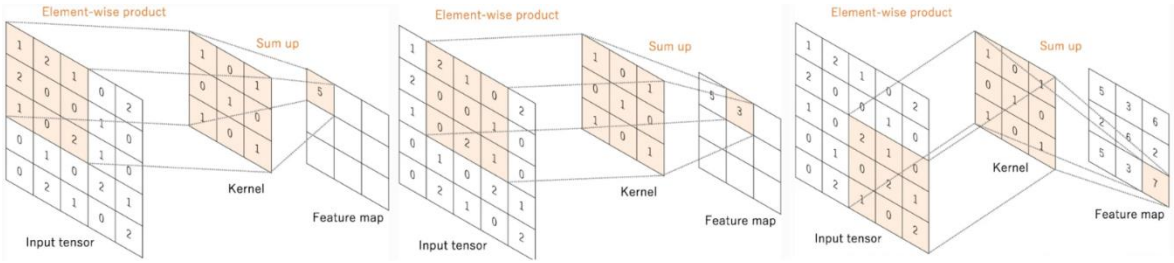



Figura 2.9. Representação de operação convolucional. Adaptado de Yamashita *et al.* (2018).

A distância entre duas posições sucessivas do kernel é denominada *stride*, que também define a operação convolucional. Geralmente, um *stride* de 1 é o mais efetivo, mas valores maiores podem ser usados com o intuito de diminuir o tamanho do mapa de recursos. A operação convolucional reduz as dimensões do mapa de recursos, em comparação ao tensor original. Para evitar esta redução, e assim possibilitar o uso de um maior número de estágios, a técnica conhecida como *padding* é utilizada, consistindo na adição de linhas e colunas contendo zeros a cada lado do tensor, mantendo as dimensões originais após a operação convolucional. Além de manter as dimensões, o uso de *padding* garante que os dados nas pontas do tensor não tenham sua importância menosprezada.

O processo de treinar uma CNN, em relação as camadas convolucionais, versa sobre identificar quais os valores dos kernels que minimizam a função objetivo em questão. Os kernels são os únicos parâmetros aprendidos automaticamente durante o processo de treinamento da camada convolucional. O número de kernels, seu tamanho, o *stride* e o número de linhas/colunas adicionadas com *padding* são hiperparâmetros definidos pelo usuário (YAMASHITA et al., 2018).

#### *2.3.1.2 Camadas pooling*

As características extraídas pela camada convolucional podem ocorrer em diferentes partes do tensor. Sua localização especifica se torna menos importante, desde que mantida a sua posição relativa as outras características extraídas. Assim, camadas *pooling* resumem informação similar nas redondezas da região e expressam como saída apenas a resposta dominante, diminuindo o tamanho do mapa de recursos (LEE; GALLAGHER; TU, 2015). Os principais exemplos desta camada são máximos (apenas o maior valor é mantido) e medios (o valor médio é mantido), mas outras abordagens podem ser utilizadas (HE et al., 2015). Neste caso, nenhum parâmetro é aprendido durante o treinamento da rede, e o tamanho, *stride* e *padding* da camada são hiperparâmetros. A Figura 2.10 representa uma camada de máximo *pooling*, com kernel [2 x 2], *stride* 2 e sem *padding*.

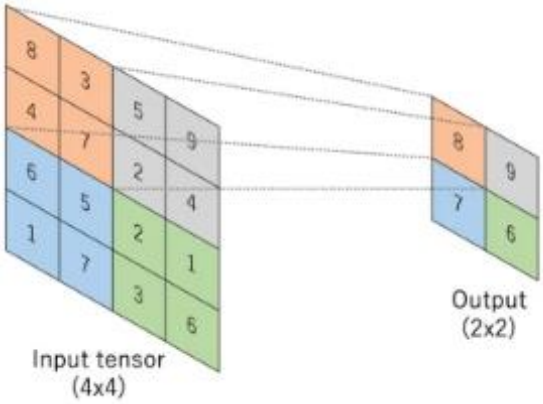



Figura 2.10. Camada max *pooling* com kernel [2x2], *stride* 2 e sem *padding*. Adaptado de Yamashita *et al.* (2018).

### *2.3.2 Normalização em lote - Batch normalization*

Sugerido por Ioffe e Szegdy (2015), a normalização em lote visa resolver o problema do deslocamento de covariância interna entre mapas de recursos. A distribuição de valores no mapa de recursos é unificada levando a média a zero e a variância unitária. O fluxo do gradiente é suavizado, aprimorando a capacidade de generalização da rede.

### *2.3.3 Evolução da arquitetura CNN*

O estudo realizado por Khan *et al*. (2020) apresenta uma extensa compilação das diversas evoluções na arquitetura das CNN ao longo dos últimos 40 anos. A Figura 2.11 serve como resumo visual dos principais pontos que marcaram essa evolução, e a nomenclatura das redes mais significativas de cada período.

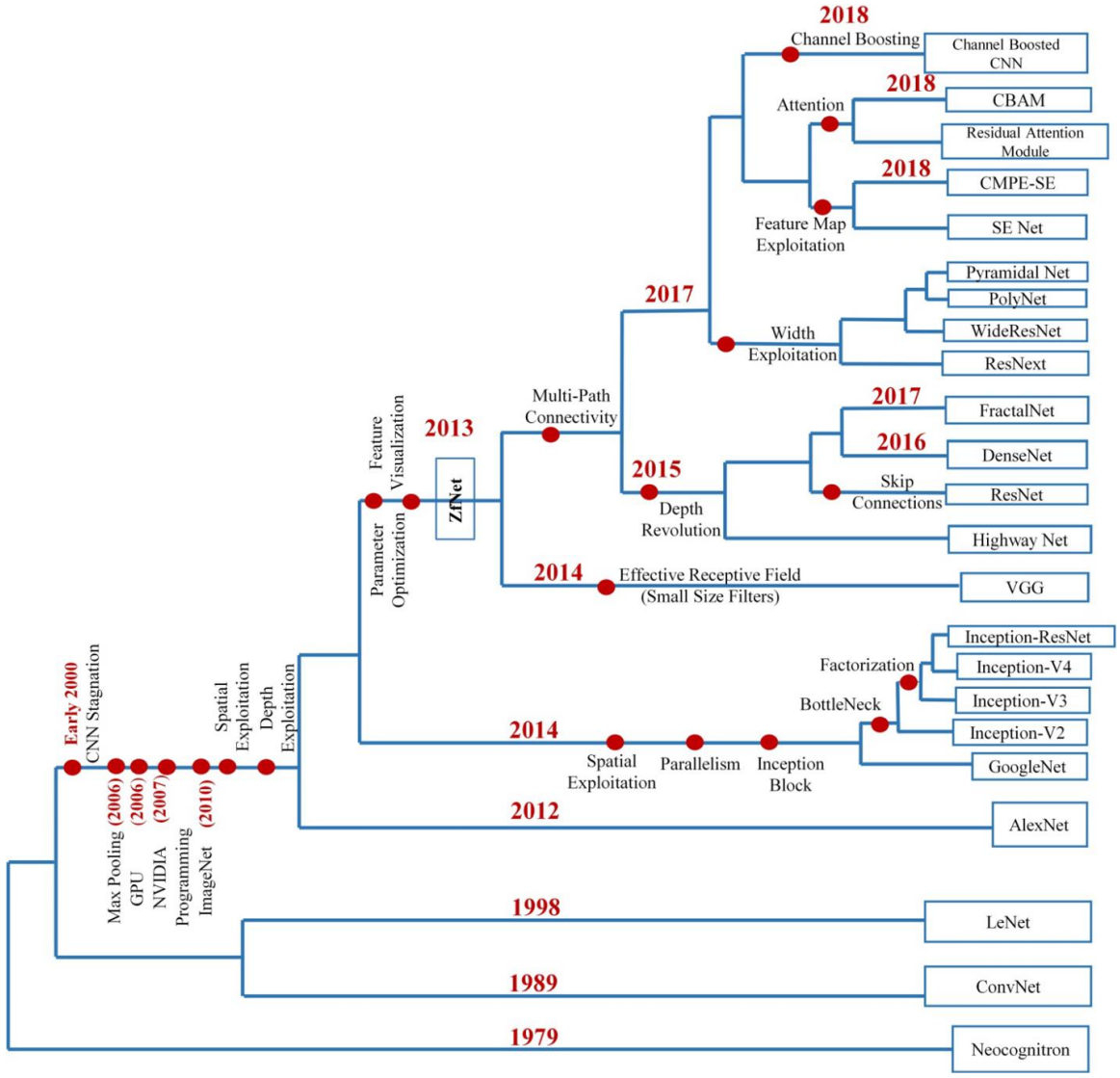



Figura 2.11. Histórico evolutivo das CNN profundas, com destaque para as inovações nas arquiteturas. Fonte: (KHAN et al., 2020).

Apesar de majoritariamente utilizadas para tratamento de imagens, redes convolucionais também podem ser aplicadas em estudos de quimiometria. Rutherford *et al*. (2020) utilizou uma CNN para interpretar matrizes excitação-emissão (EEM) de fluorescência, a fim de identificar a fonte emissora de material particulado proveniente de combustão. Por sua vez, Itakura *et al.* (2018) utilizou uma CNN regressora para estimar maturidade de citros, também com base em EEM de fluorescência. A natureza bidimensional das matrizes espectrais abre a possibilidade para aplicação de classificação ou regressão utilizando CNN, área ainda pouco explorada industrialmente.

### *2.3.4 Redes Autoencoder - AE*

Autoencoder são uma técnica de aprendizado não supervisionado que pretende capturar as características do sistema em uma camada interna, que age como extrator de padrões dos dados (WETZEL, 2017). AE podem ser apresentados como uma generalização não linear da análise de componentes principais (PCA), utilizando uma rede neural simétrica para primeiramente codificar os dados de entrada em um novo espaço latente,

e, posteriormente, decodificar este espaço, buscando a reconstrução do espaço original. Restrições são impostas para que a rede não simplesmente se torne uma identidade, repetindo diretamente as entradas nas saídas (CHOLLET, 2015). A Figura 2.12 apresenta um AE genérico do tipo *bottleneck*, com redução da dimensão do espaço original.

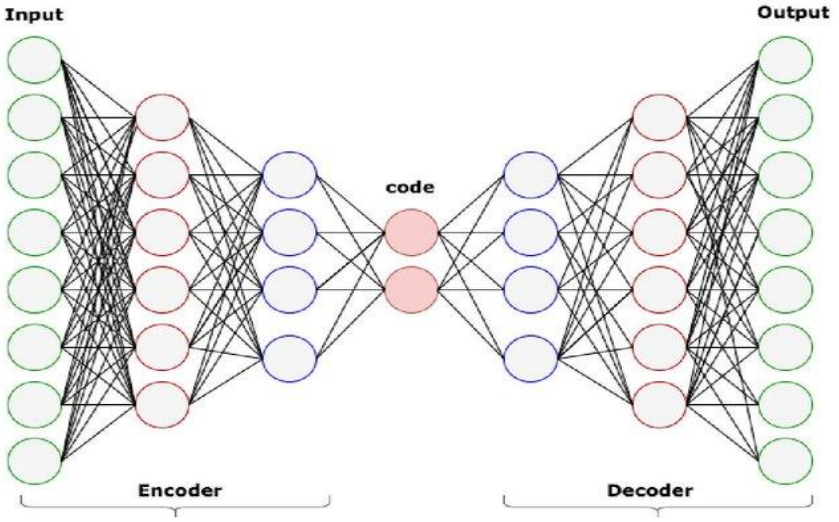



Figura 2.12. Esquema genérico da arquitetura de um AE. Fonte: (BAHI; BATOUCHE, 2018).

Diversos trabalhos na literatura comparam o uso de PCA e AE, geralmente explicitando as vantagens das redes em capturar padrões não lineares nos dados, identificando características no espaço latente que a PCA não consegue capturar (HINTON; SALAKHUTDINOV, 2006; JES et al., 2019; MANNING-DAHAN, 2017). Esta diferença pode ser vista na Figura 2.13, comparando a projeção espacial dos dois primeiros componentes principais – PCA (A) e do espaço latente de 2 neurônios da rede AE (B), para o conjunto dígitos escritos à mão, MNIST. Assim como no caso da PCA, o espaço latente gerado pelo AE pode ser utilizado como base para modelos de regressão (HINTON; SALAKHUTDINOV, 2006).

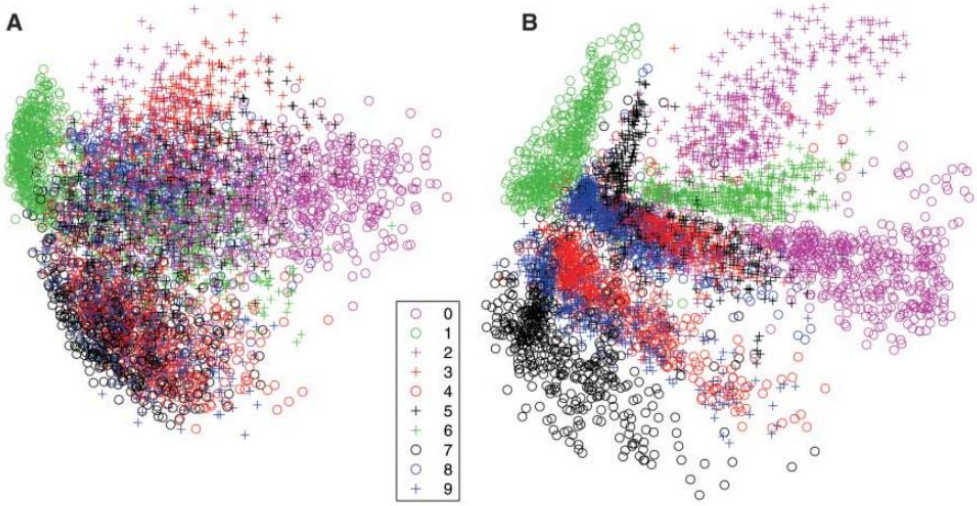



Figura 2.13. Projeção espacial dos dois primeiros componentes principais – PCA (A) e do espaço latente de 2 neurônios da rede AE (B), para o conjunto dígitos escritos a mão, MNIST. Fonte: (HINTON; SALAKHUTDINOV, 2006).

A construção de um AE segue os mesmos princípios básicos de outras redes neurais, e pode ser resumido no esquema apresentado na Figura 2.14.

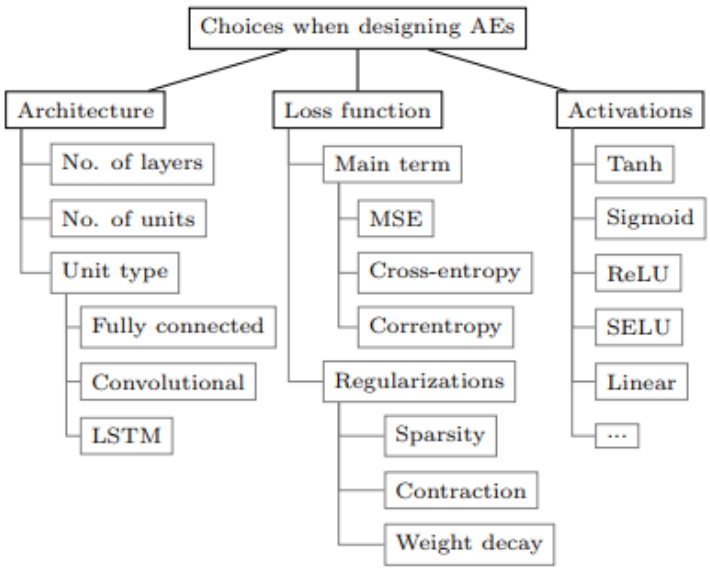



Figura 2.14. Resumo das escolhas necessárias para o design de um AE. Fonte: (CHARTE et al., 2018).

Além da clássica rede *feedforward* apresentada, diversas arquiteturas de redes AE são propostas na literatura para resolver problemas específicos, como:

- *Denoising* AE – DAE (VINCENT et al., 2008): adaptação desenvolvida para criar uma rede que fosse robusta a dados corrompidos. A DAE aprende a gerar padrões robustos das entradas reconstruindo amostras parcialmente destruídas. A estrutura e parâmetros são idênticos a de um AE comum, mas, durante a fase de treinamento da rede, erros estocásticos corrompem as entradas, enquanto o erro de reconstrução a ser minimizado ainda considera a base de dados original. Assim, a rede é treinada para predizer os valores falhos, e reconstruir os dados originais de maneira a não depender destas entradas.

- *Convolutional AE – CAE* (MASCI et al., 2011): AE padrões não explicitamente consideram a estrutura bidimensional quando processam os dados. CAE resolvem este problema aplicando camadas convolucionais ao invés de camadas densas. Em alguns casos, CAE são utilizadas como ponto de partida para inicializar CNN.

- *LSTM AE* (SRIVASTAVA; MANSIMOV; SALAKHUTDINOV, 2015): AE básicos não são planejados para modelar dados sequenciais. O uso de unidades LSTM (*Long-Short-Term Memory*) como codificadores e decodificadores da rede resolve este desafio. O codificador comprime a sequência em uma representação de tamanho fixo, e o decodificador treina para extrair a sequência original na ordem inversa. Esta abordagem é especialmente útil para dados sequenciais e grandes, como arquivos de vídeo.

- *Variational AE* – VAE (KINGMA; WELLING, 2014): Este tipo de AE aplica uma abordagem variacional Bayesiana para codificação. Se assume que existe uma variável aleatória latente não observada y, que por um processo randômico leva

as observações, x. O objetivo é aproximar a distribuição das variáveis latentes dadas as observações. VAE substituem as funções determinísticas no codificador e decodificador por mapeamento estocástico, e calculam a função objetivo em virtude da função densidade das variáveis randômicas. A principal vantagem deste tipo de rede é que é possível gerar novas amostras a partir da distribuição aprendida, e são muito aplicadas em casos que envolvem a necessidade de gerar novas instâncias (DOSOVITSKIY; BROX, 2016).

As principais aplicações das diversas abordagens de AE são a classificação (transformam os dados de entrada para obter melhores performances em classificadores); compressão de dados (AE treinada com um tipo especifico de dados para aprender formas eficientes de compressão); *hashing* (resumem os dados de entrada em vetores binários para acelerar buscas); e visualização (projetam os dados em duas ou três dimensões para representação gráfica).

Uma característica intrínseca das AE é seu potencial para detecção de padrões anormais nos dados. Como a rede é treinada para fielmente reconstruir o padrão aprendido dos dados de calibração, amostras anormais se destacam na base de dados, apresentando erros de reconstrução. Sakurada e Yairi (2014) comparam o potencial de um AE para detectar dados anormais em um sistema artificial e em dados reais de telemetria de aeronaves, comparando com outras abordagens utilizando PCA e kernel PCA. Em ambos os casos, a performance do AE foi superior as demais metodologias. Zou *et al.* (2019) e Misra *et al*. (2020) propõem o uso de AE para detecção de fraudes em cartão de crédito. Amarbayasgalan *et al*. (2018) propõe uma nova metodologia baseada em DAE e clusterização para identificação de dados anômalos em 20 bases de dados de referência. Os autores comparam a metodologia proposta com outras 12 metodologias, baseadas em *Support Vector Machines* e PCA. A metodologia sugerida é superior aos algoritmos estado-da-arte para classificação de 9 das bases avaliadas.

## 2.4 Espectroscopia

A espectroscopia estuda a interação da matéria com radiação eletromagnética. A radiação pode ser absorvida, transmitida ou dispersa pela matéria. Métodos espectroscópicos são aqueles que analisam a quantidade de radiação absorvida ou produzida pelo meio. Tais métodos são classificados de acordo com a região do espectro eletromagnético envolvido e a reação da matéria a esta energia (SETTLE, 1997).

Dentre os métodos óticos, se destacam a espectroscopia por fluorescência e a espectroscopia vibracional (espectroscopia no infravermelho próximo e médio (YADAV; YADAV, 2005) e a espectroscopia Raman (JONES et al., 2019)). A espectroscopia de absorção no infravermelho estuda a interação de radiação nas frequências especificas capazes de alterar o nível vibracional ou rotacional das moléculas. O deslocamento dos átomos em vibração altera o momento dipolar da molécula, e causa estiramento das ligações. Como a energia de cada nível quântico vibracional é especifica, a análise da absorção de diversas frequências permite determinar a composição química da amostra. A espectroscopia Raman, por sua vez, analisa o espalhamento inelástico da luz pela matéria. Como este espalhamento depende da estrutura da molécula e de seus níveis de energia vibracional, a espectroscopia Raman consegue detectar o tipo de estrutura, de ligações e outras características físico-químicas do composto. A Tabela 2.1 apresenta algumas características e aplicações da espectroscopia vibracional.

Tabela 2.1. Características e aplicações de técnicas de espectroscopia vibracional.

| Técnica | Características | Aplicações |
|---|---|---|
| Espectroscopia de infravermelho médio (MIR) | Espectroscopia de absorção<br><br>Modos fundamentais de vibração (alta intensidade) | Ciência dos materiais (KISCHKAT et al., 2012; PALUSZKIEWICZ et al., 2011), agricultura (ELLEN MACARTHUR FOUNDATION, 2013; SORIANO-DISLA et al., 2014), ciência de alimentos (KAROUI; DE BAERDEMAEKER, 2007; SINELLI et al., 2008), indústria farmacêutica (BRITTAIN, 2018) |
| Espectroscopia de infravermelho próximo (NIR) | Espectroscopia de absorção<br><br>Sobretons e bandas de combinação (mais fracas) | Ciências agrícolas (WILSON; ZHANG; KOVACS, 2014), pecuária (VALENTI et al., 2013), biotecnologia (YU et al., 2014), combustíveis (CRAMER et al., 2009), química fina (SHINZAWA et al., 2012) |
| Espectroscopia Raman | Espectroscopia de espalhamento<br><br>Composição molecular e estrutural | Cristalografia (CAREY, 2014), ciência dos materiais (DRESSELHAUS; JORIO; SAITO, 2010), ciências biológicas (RYGULA et al., 2013), ciência de alimentos (KRIEG, 2014; REID; O'DONNELL; DOWNEY, 2006), indústria farmacêutica (ŠAŠIĆ, 2007) |

A luminescência é uma família de processos onde moléculas suscetíveis emitem luz a partir de estados eletrônicos excitados, seja por meio físico (absorção de radiação), mecânico (fricção) ou químico (reação). Quando a geração de luminescência parte da excitação de uma molécula por fótons, o fenômeno se determina fotoluminescência, e é formalmente dividido entre fluorescência e fosforescência, dependendo do tipo da configuração eletrônica do estado excitado e do caminho de emissão de luz. O fenômeno da fluorescência é aquele observado em estados excitados do tipo singleto (o spin do elétron presente no orbital excitado se encontra pareado com o spin do elétron do estado fundamental). No caso da fosforescência, o spin do elétron excitado se encontra desemparelhado, permanecendo paralelo, em um estado denominado tripleto, mais improvável de acontecer (SKOOG; HOLLER; CROUCH, 2009). O Diagrama de Jablonski, apresentado na Figura 2.15, ilustra os processos de mudança de níveis energéticos entre absorção e emissão de luz, e serve como base para o entendimento dos fenômenos de fotoluminescência.

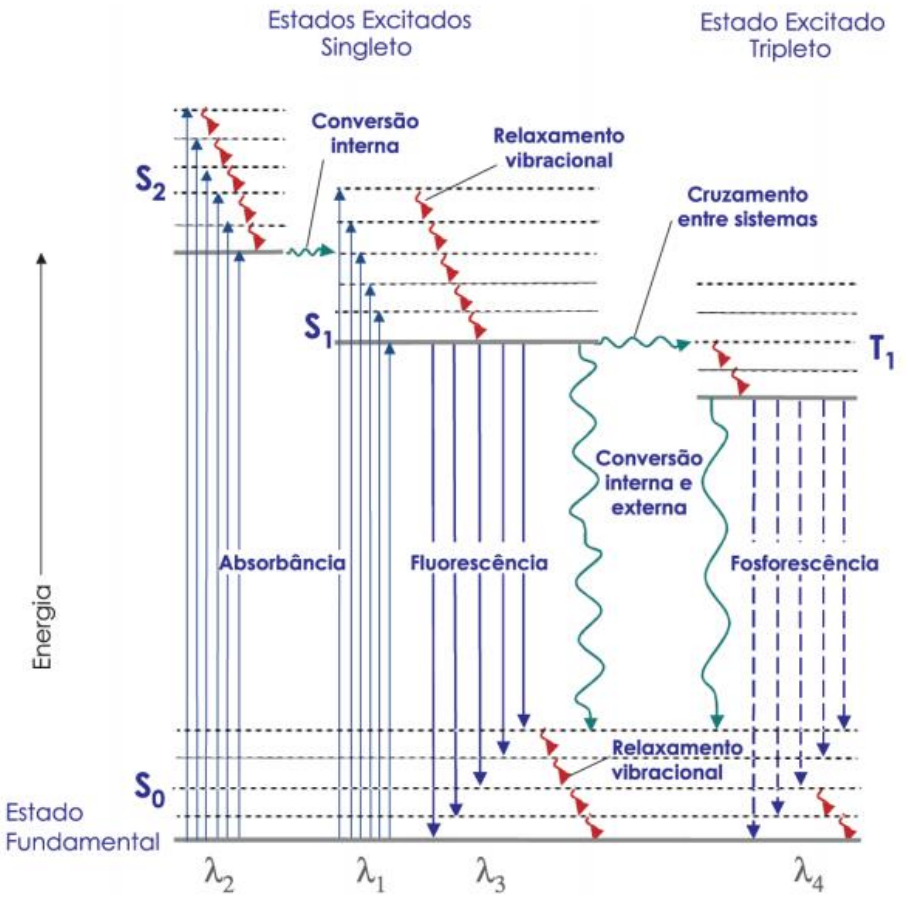



Figura 2.15. Diagrama de Jablonski. $S_0$ representa o estado eletrônico fundamental, $S_1$, e $T_1$ são os estados excitados singleto e tripleto. $S_2$ é um segundo estado excitado singleto. As linhas horizontais pontilhadas são os vários níveis de energia vibracional dos estados. As setas retas representam os processos envolvendo fótons, e as setas onduladas representam transições não-radioativas. Fonte: (SKOOG; HOLLER; CROUCH, 2009).

A análise das relações entre energias presentes no Diagrama de Jablonski explicam grande parte das características do processo de fluorescência. O processo é governado majoritariamente por três eventos: primeiramente, uma molécula suscetível é excitada por fótons. A excitação ocorre do estado fundamental $S_0$ para estados singletos $S_1$ e $S_2$, em diferentes níveis vibracionais de energia. A energia é geralmente dissipada através de relaxamento vibracional, levando o sistema ao menor nível energético do estado excitado $S_1$. Por fim, fótons são emitidos pela molécula, e a mesma retorna ao estado fundamental. Muitas vezes, ela retorna para níveis vibracionais de maior energia no estado fundamental, e rapidamente relaxa para o estado de menor energia.

### *2.4.1 Absorção, Emissão e Deslocamento de Stokes*

Para um típico fluoróforo, a irradiação por um amplo espectro de comprimentos de ondas irá gerar toda uma gama de transições permitidas, que irão povoar os vários níveis de energia vibracional dos estados excitados. Algumas destas transições terão maior probabilidade de acontecer do que outras, e, quando combinadas, constituem o espectro de absorção da molécula. A probabilidade de uma transição ocorrer entre o estado fundamental $S_0$ para um estado excitado $S_1$ depende da similaridade entre os níveis de energia vibracional e rotacional dos estados. O comprimento de onda de máxima absorção representa a mais provável separação entre o estado fundamental e um nível vibracional permitido no estado excitado.

Após a absorção de um fóton, diversos processos com diferentes probabilidades ocorrem, sendo os mais comuns deles a conversão interna e o relaxamento vibracional, levando a molécula ao mais baixo nível de energia vibracional do estado excitado $S_1$. Neste momento, a molécula pode emitir um fóton, no processo que se conhece como fluorescência. As análises dos diferentes comprimentos de onda emitidos por uma molécula ao ser excitada dão origem ao espectro de emissão. Como é raro que uma molécula emita fótons a partir de estados excitados superiores ao $S_1$, ocorrendo rápido relaxamento vibracional para o menor nível de energia do estado excitado, o espectro de emissão de uma molécula é geralmente independente do comprimento de onda de excitação. Este fenômeno pode ser visto na Figura 2.15, acompanhado as setas vermelhas onduladas, e as setas retas de fluorescência. Ao retornar para o estado fundamental, vários níveis de energia vibracional podem ser povoados.

Assim como para absorção, a probabilidade de um elétron em um estado excitado retornar para um nível de energia vibracional específico no estado fundamental é proporcional à sobreposição entre os níveis de energia nos respectivos estados. Desta forma, o espectro de emissão é geralmente uma imagem espelhada do espectro de absorção do estado $S_0$ para o estado $S_1$. De fato, a probabilidade de um elétron retornar a um nível especifico de energia vibracional no estado fundamental é similar a probabilidade daquele elétron pertencer aquele nível antes da excitação. Este conceito é conhecido como Regra da Imagem Espelhada.

A Figura 2.16 apresenta os espectros de absorção e de emissão de duas moléculas distintas. Os espectros do perileno (superior) seguem a regra da imagem espelhada. Neste caso, as distâncias entre os níveis de energia vibracional são consideráveis e isso se reflete nos múltiplos picos do espectro. Os espectros da quinina (inferior) não seguem a regra da imagem espelhada. Neste caso, o primeiro pico de absorção (315 nm) é referente a uma transição entre os estados $S_0$ e $S_2$. Como a energia interna é relaxada e a molécula decresce ao menor nível vibracional de $S_1$ antes de emitir fótons, o espectro de emissão não apresenta este pico, e é uma imagem espelhada das transições $S_0 – S_1$. No caso da quinina, as distâncias entre os níveis de energia vibracional são muito próximas e o espectros se apresentam na forma de bandas mais largas.

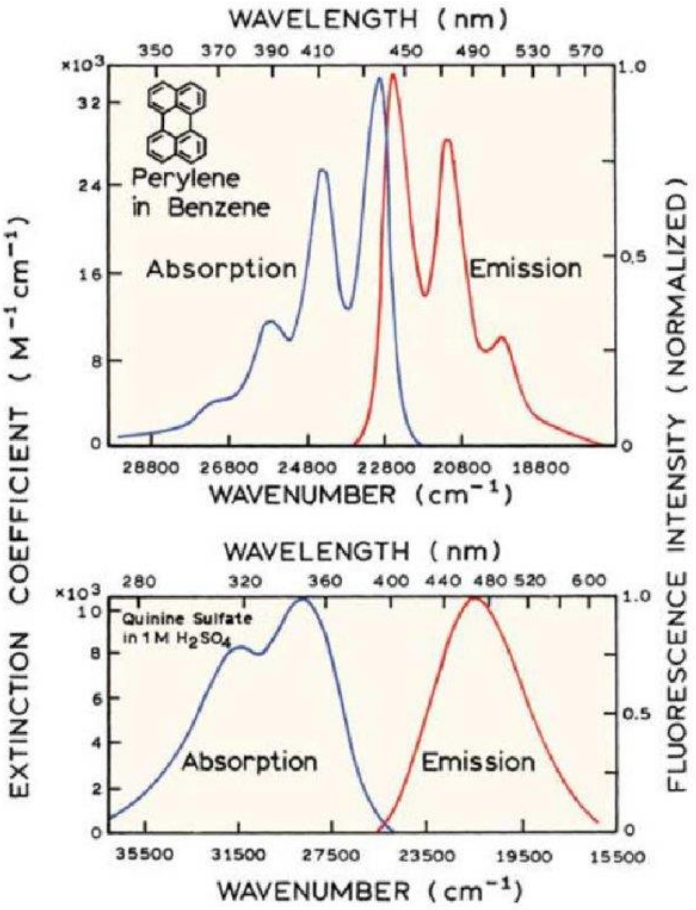



Figura 2.16. Espectro de absorção e emissão de fluorescência do perileno (superior) e da quinina (inferior). O Deslocamento de Stokes pode ser visto claramente em ambas moléculas. O perileno segue a regra da Imagem Espelhada, mas o mesmo não ocorre para a quinina. Fonte: (LAKOWICZ, 2006).

A energia associada às transições de emissão são tipicamente menores que aquelas ligadas a absorção, resultando na emissão de fótons com menor energia (e maiores comprimentos de onda). Este fenômeno é conhecido como Deslocamento de Stokes, e ocorre para virtualmente todos os fluoróforos. A principal causa do efeito é a perda de energia por relaxamento dos elétrons excitados para o menor nível de energia vibracional do estado excitado $S_1$. Além disso, a emissão geralmente acontece em transições para níveis de energia vibracional superiores do estado fundamental, e o equilíbrio térmico é novamente alcançando pelo relaxamento vibracional da energia excedente. A existência do Deslocamento de Stokes é essencial para extrema sensibilidade das medidas de fluorescência. O deslocamento dos comprimentos de onda entre absorção e emissão faz com que seja possível o uso de filtros óticos precisos, capazes de bloquear a luz de excitação da fonte emissora, e assim o sinal de emissão pode ser observado em um sistema de baixo ruído.

### *2.4.2 Matriz Excitação – Emissão e Espectrofluorímetro*

Outra forma comum de estudar os efeitos da fluorescência é através da matriz excitação – emissão (*Excitation-Emission Matrix* – EEM), onde os comprimentos de onda de excitação e emissão são varridos em ambos os módulos e apresentados na forma de um mapa tridimensional $\lambda_{\text{excitação}}\ x\ \lambda_{emissão}\ x\ intensidade\ de\ fluorescência$. Comparado com outros tipos de espectros por fluorescência, os resultados na forma EEM apresentam

uma detecção multidimensional muito mais compreensiva da fluorescência, especialmente para amostras complexas (RUTHERFORD et al., 2020). Um típico EEM pode ser visto na Figura 2.17, tanto na forma tridimensional, quando na forma de curvas de contorno. Uma grande variedade de informações pode ser extraídas das EEM, como intensidade, localização e distribuição dos picos, informação decorrente da decomposição espectral e informação relacionada a energia dos fótons. Dada a natureza dos dados, são geralmente necessárias análises quimiométricas complexas para pré-tratar e extrair informação útil das matrizes, nas quais se destacam as técnicas de PCA/PLS (FAASSEN; HITZMANN, 2015; RANZAN et al., 2017), PARAFAC (BAHRAM et al., 2006; GUIMET et al., 2004; RÍOS-REINA et al., 2017; SILVA et al., 2019), meta-heurísticas de seleção de variáveis (ASSAWAJARUWAN; REINALTER; HITZMANN, 2017; RANZAN et al., 2015) e redes neurais convolucionais (ITAKURA et al., 2018; RUTHERFORD et al., 2020).

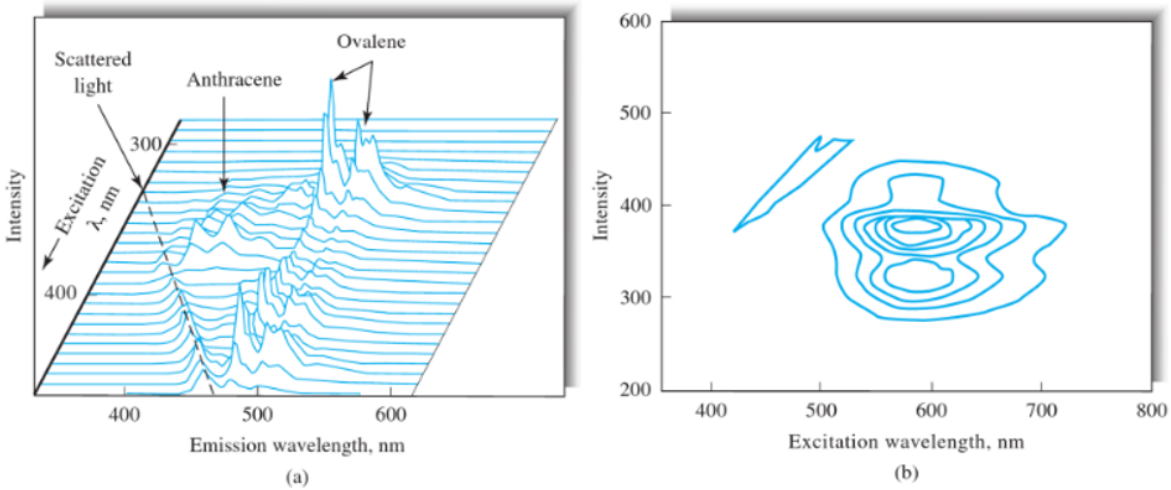



Figura 2.17. Matrizes excitação – emissão de fluorescência para uma mistura de antraceno e ovaleno (a), apresentada na forma tridimensional, e para 8-hidroxibenzopireno (b), apresentada como curvas de contorno. Fonte: (SKOOG; HOLLER; CROUCH, 2009).

O equipamento utilizado para obtenção de EEM de fluorescência se denomina espectrofluorímetro. Sua construção permite a obtenção tanto de espectros de excitação (comprimento de onda de emissão constante), quanto espectros de emissão (comprimento de onda de excitação constante), por meio do uso de dois monocromadores. Assim, é factível a construção dos mapas espectrais variando ambos os módulos sucessivamente. A Figura 2.18 apresenta a estrutura genérica de um espectrofluorímetro. A radiação oriunda da fonte passa por um monocromador para seleção do comprimento de onda de excitação. O feixe é então dividido em duas direções, uma que passará pela amostra, e outra que será atenuada e direcionada à fotomultiplicadora de referência. A radiação fluorescente da amostra, por sua vez, passa por um monocromador para seleção do comprimento de onda de emissão e é detectada pela fotomultiplicadora de emissão. A emissão de fluorescência da amostra se dá em todas as direções, mas é geralmente medida em ângulo reto com o feixe de excitação, para atenuar interferências. A eficiência dos monocromadores determinará a resolução do equipamento.

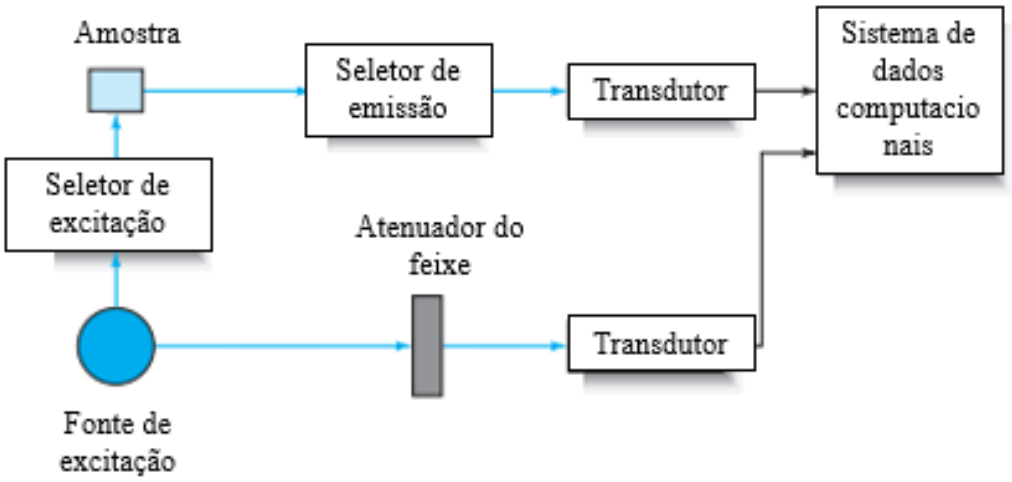



Figura 2.18. Representação genérica da estrutura de um espectrofluorímetro. Fonte: (SKOOG; HOLLER; CROUCH, 2009).

### *2.4.3 Supressão de Fluorescência - Quenching*

A intensidade de fluorescência pode ser suprimida por uma ampla gama de processos, denominados de *quenching*. No *quenching* colisional, uma molécula excitada pode ser desativada ao colidir com outras moléculas da solução, denominados *quenchers*. Desta forma, ocorre troca de energia e o fluoróforo deixa de emitir fótons. Quanto maior a temperatura do meio, maiores as chances destas colisões acontecem. Exemplos de *quenchers* comuns incluem oxigênio, halogênios, átomos pesados, aminas e moléculas eletro-deficientes. A fluorescência pode também ser suprimida pela formação de complexos não fluorescentes entre o fluoróforo e um *quencher*. Alguns fluoróforos também podem formar complexos consigo mesmos, como a formação de excímeros ou polímeros, alterando sua fluorescência a depender da concentração molecular. O pH também pode influenciar a fluorescência do meio, alterando as relações energéticas dos elétrons. Por fim, diversos mecanismos não moleculares podem causar *quenching*, como efeitos de filtro interno (absorção da emissão pelo próprio meio) e efeitos de barreira física, com atenuação da luz incidente (LAKOWICZ, 2006).

## 2.5 Referências

ABDULKADER, M. M. S.; GAJPAL, Y.; ELMEKKAWY, T. Y. Hybridized ant colony algorithm for the Multi Compartment Vehicle Routing Problem. **Applied Soft Computing**, v. 37, p. 196–203, 2015.

AMARBAYASGALAN, T.; JARGALSAIKHAN, B.; RYU, K. Unsupervised Novelty Detection Using Deep Autoencoders with Density Based Clustering. **Applied Sciences**, v. 8, 2018.

ASSAWAJARUWAN, S.; REINALTER, J.; HITZMANN, B. Comparison of methods for wavelength combination selection from multi-wavelength fluorescence spectra for on-line monitoring of yeast cultivations. **Analytical and Bioanalytical Chemistry**, v. 409, n. 3, p. 707–717, 2017.

BAHI, M.; BATOUCHE, M. **Deep semi-supervised learning for DTI prediction using large datasets and H2O-spark platform**. 2018 International Conference on Intelligent Systems and Computer Vision, ISCV 2018. **Anais**...Institute of Electrical and Electronics Engineers Inc., 3 maio 2018

BAHRAM, M. et al. Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation. **Journal of Chemometrics**, v. 20, n. 3–4, p. 99–105, 1

mar. 2006.

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, n. 1, p. 63–72, 2011.

BALSEIRO, S.; LOISEAU, I.; RAMONET, J. An Ant Colony algorithm hybridized with insertion heuristics for the Time Dependent Vehicle Routing Problem with Time Windows. **Computers & OR**, v. 38, p. 954–966, 2011.

BASHEER, I. A.; HAJMEER, M. Artificial neural networks: fundamentals, computing, design, and application. **Journal of Microbiological Methods**, v. 43, n. 1, p. 3–31, 2000.

BHATTACHARYYA, S. **Hybrid Metaheuristics**. [s.l.] WORLD SCIENTIFIC, 2018. v. 84

BLUM, C. et al. **Hybrid Metaheuristics: An Emerging Approach to Optimization**. [s.l: s.n.].

BLUM, C. et al. Hybrid metaheuristics in combinatorial optimization: A survey. **Applied Soft Computing**, v. 11, n. 6, p. 4135–4151, 2011.

BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A. A review of feature selection methods on synthetic data. **Knowledge and Information Systems**, v. 34, n. 3, p. 483–519, 2013.

BRITTAIN, H. G. Mid-Infrared Spectroscopy of Pharmaceutical Solids. In: **Profiles of Drug Substances, Excipients and Related Methodology**. [s.l.] Academic Press Inc., 2018. v. 43p. 321–358.

BRUCKER, P. NP-Complete operations research problems and approximation algorithms. **Zeitschrift für Operations Research**, v. 23, n. 3, p. 73–94, 1 jun. 1979.

CAMACHO, J.; PICÓ, J.; FERRER, A. Data understanding with PCA: Structural and Variance Information plots. **Chemometrics and Intelligent Laboratory Systems**, v. 100, n. 1, p. 48–56, 2010.

CAREY, P. Raman Crystallography, the Missing Link Between Biochemical Reactions and Crystallography. In: [s.l.] Springer, Dordrecht, 2014. p. 13–24.

CHANDRA MOHAN, B.; BASKARAN, R. A survey: Ant Colony Optimization based recent research and implementation on several engineering domain. **Expert Systems with Applications**, v. 39, n. 4, p. 4618–4627, 2012.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers and Electrical Engineering**, v. 40, n. 1, p. 16–28, 1 jan. 2014.

CHARTE, D. et al. A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. **Information Fusion**, v. 44, n. December 2017, p. 78–96, 2018.

CHI, Q. et al. A model predictive control approach with relevant identification in dynamic PLS framework. **Control Engineering Practice**, v. 22, n. 0, p. 181–193, 2014.

CHOLLET, F. **Deep Learning With Python**. [s.l: s.n.]. v. 1

CIREŞAN, D. et al. Multi-column deep neural network for traffic sign classification. **Neural Networks**, v. 32, p. 333–338, ago. 2012.

CRAMER, J. A. et al. Ultra-low sulfur diesel classification with near-infrared spectroscopy and partial least squares. **Energy and Fuels**, v. 23, n. 2, p. 1132–1133, 2009.

CYBENKO, G. Approximation by superpositions of a sigmoidal function. **Mathematics of Control, Signals, and Systems**, v. 2, n. 4, p. 303–314, dez. 1989.

DORIGO, M.; BLUM, C. Ant colony optimization theory: A survey. **Theoretical Computer Science**, v. 344, n. 2–3, p. 243–278, 2005.

DORIGO, M.; GAMBARDELLA, L. M. Ant colony system: A cooperative learning approach to the traveling salesman problem. **IEEE Transactions on Evolutionary Computation**, v. 1, n. 1, p. 53–66, 1997.

DOSOVITSKIY, A.; BROX, T. Generating Images with Perceptual Similarity Metrics based on Deep Networks. **Advances in Neural Information Processing Systems**, p. 658–666, 2016.

DRESSELHAUS, M. S.; JORIO, A.; SAITO, R. Characterizing graphene, graphite, and carbon nanotubes by Raman spectroscopy. **Annual Review of Condensed Matter Physics**, 2010.

EFRON, B. et al. Least angle regression. **The Annals of Statistics**, v. 32, n. 2, p. 407–499, 2004.

ELLEN MACARTHUR FOUNDATION. **Towards the circular economy. Economic and business rationale for an accelerated transition.**

FAASSEN, S. M.; HITZMANN, B. Fluorescence Spectroscopy and Chemometric Modeling for Bioprocess Monitoring. p. 10271–10291, 2015.

FILZMOSER, P.; TODOROV, V. Review of robust multivariate statistical methods in high dimension. **Analytica Chimica Acta**, v. 705, n. 1–2, p. 2–14, 2011.

GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. **Analytica Chimica Acta**, v. 185, n. 0, p. 1–17, 1986.

GLOVER, F. Tabu Search—Part I. **ORSA Journal on Computing**, v. 1, n. 3, p. 190–206, 1 ago. 1989.

GLOVER, F. Tabu Search—Part II. **ORSA Journal on Computing**, v. 2, n. 1, p. 4–32, 1 fev. 1990.

GODOY, J. L.; VEGA, J. R.; MARCHETTI, J. L. Relationships between PCA and PLS-regression. **Chemometrics and Intelligent Laboratory Systems**, v. 130, p. 182–191, 15 jan. 2014.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization and Machine Learning**. Boston: Kluwer Academic Publishers, 1989.

GUIDOTTI, R. et al. A Survey of Methods for Explaining Black Box Models. **ACM Computing Surveys**, v. 51, 2018.

GUIMET, F. et al. Application of unfold principal component analysis and parallel factor analysis to the exploratory analysis of olive oils by means of excitation–emission matrix fluorescence spectroscopy. **Analytica Chimica Acta**, v. 515, n. 1, p. 75–85, 2004.

GUYON, I. et al. Gene Selection for Cancer Classification using Support Vector Machines. **Machine Learning**, v. 46, n. 1, p. 389–422, 2002.

HASTIE, TREVOR, TIBSHIRANI, ROBERT, FRIEDMAN, J. **The Elements of Statistical Learning The Elements of Statistical LearningData Mining, Inference, and Prediction, Second Edition**. [s.l: s.n.].

HE, K. et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 37, n. 9, p. 1904–1916, 1 set. 2015.

HEIYANTHUDUWAGE, M. A.; MOUNOURY, S.; KOVACEVIC, A. **Performance prediction methods for screw compressors**. Institution of Mechanical Engineers - 7th International Conference on Compressors and Their Systems 2011. **Anais**...Woodhead Publishing Limited, 1 jan. 2011

HEMMATEENEJAD, B.; MIRI, R.; ELYASI, M. A segmented principal component analysis—

regression approach to QSAR study of peptides. **Journal of Theoretical Biology**, v. 305, n. 0, p. 37–44, 2012.

HINTON, G. E.; SALAKHUTDINOV, R. R. Reducing the dimensionality of data with neural networks. **Science**, v. 313, n. 5786, p. 504–507, 28 jul. 2006.

HOERL, A. E.; KENNARD, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. **Technometrics**, v. 12, n. 1, p. 55–67, 1970.

HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Multilayer feedforward networks are universal approximators. **Neural Networks**, v. 2, n. 5, p. 359–366, 1 jan. 1989.

HU, L. et al. Vis-NIR spectroscopy Combined with Wavelengths Selection by PSO Optimization Algorithm for Simultaneous Determination of Four Quality Parameters and Classification of Soy Sauce. **Food Analytical Methods**, v. 12, n. 3, p. 633–643, 2019.

IOFFE, S.; SZEGEDY, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015.

ITAKURA, K. et al. Estimation of Citrus Maturity with Florescence Spectroscopy Using Deep Learning. **Horticulturae**, v. 5, n. 1, p. 2, 26 dez. 2018.

JACKSON, J. E. **A User's Guide to Principal Components**. Chichester: Wiley, 1991.

JES, F. et al. Deep Convolutional Autoencoders vs PCA in a Highly-Unbalanced Parkinson ' s Disease Dataset : A DaTSCAN Study. **Springer Nature**, p. 47–56, 2019.

JOHNSON, B.; XIE, Z. Classifying a high resolution image of an urban area using super-object information. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 83, p. 40–49, 2013.

JOLLIFFE, I. T. C. N.-S. C. I. Q. 5. J. 1986 H. L. Q. 5. J. 1986. **Principal component analysis**. New York: Springer-Verlag, 1986.

JONES, R. R. et al. **Raman Techniques: Fundamentals and FrontiersNanoscale Research Letters**Springer New York LLC, , 1 dez. 2019. Disponível em: </pmc/articles/PMC6626094/?report=abstract>. Acesso em: 18 ago. 2020

KAROUI, R.; DE BAERDEMAEKER, J. **A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy productsFood Chemistry**, 2007.

KAVZOGLU, T. Chapter 33 - Object-Oriented Random Forest for High Resolution Land Cover Mapping Using Quickbird-2 Imagery. In: SAMUI, P.; SEKHAR, S.; BALAS, V. E. (Eds.). . **Handbook of Neural Computation**. [s.l.] Academic Press, 2017. p. 607–619.

KENNEDY, J.; KENNEDY, J.; EBERHART, R. Particle swarm optimization. p. 4--1942, 1995.

KHAN, A. et al. A survey of the recent architectures of deep convolutional neural networks. **Artificial Intelligence Review**, p. 1–62, 21 abr. 2020.

KINGMA, D. P.; WELLING, M. **Auto-encoding variational bayes**. 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings. **Anais**...International Conference on Learning Representations, ICLR, 20 dez. 2014Disponível em: <https://arxiv.org/abs/1312.6114v10>. Acesso em: 7 set. 2020

KIRKPATRICK, S.; GELATT, C. D.; VECCHI, M. P. Optimization by simulated annealing. **Science**, 1983.

KISCHKAT, J. et al. Mid-infrared optical properties of thin films of aluminum oxide, titanium dioxide, silicon dioxide, aluminum nitride, and silicon nitride. **Applied Optics**, v. 51, n. 28, p. 6789–6798, 1 out. 2012.

KRIEG, T. Real-time monitoring of continuous fermentation by Raman spectroscopy. n. September, p. 30, 2014.

LAKOWICZ, J. R. C. N.-S. C. I. Q. F. L. 2006 N. E. T. M. I. T. A. O. **SEE U. H. L. Q. F. L. 2006 H. L. C. Q. F. L. 2006. **Principles of fluorescence spectroscopy**. 3rd. ed. New York: Springer, 2006.

LECUN, Y.; KAVUKCUOGLU, K.; FARABET, C. **Convolutional networks and applications in vision**. ISCAS 2010 - 2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems. **Anais**...2010

LEE, C.-Y.; GALLAGHER, P. W.; TU, Z. Generalizing Pooling Functions in Convolutional Neural Networks: Mixed, Gated, and Tree. **Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016**, p. 464–472, 29 set. 2015.

LEIJNEN, S.; VAN VEEN, F. The Neural Network Zoo †. **Proceedings 2020, Vol. 47, Page 9**, v. 47, n. 1, p. 9, 12 maio 2020.

LI, X. Y.; TIAN, P.; LEUNG, S. C. H. An ant colony optimization metaheuristic hybridized with tabu search for open vehicle routing problems. **Journal of the Operational Research Society**, v. 60, n. 7, p. 1012–1025, 2009.

LIU, X.; DENG, Z.; YANG, Y. Recent progress in semantic image segmentation. **Artificial Intelligence Review**, v. 52, n. 2, p. 1089–1106, 15 ago. 2019.

MANNING-DAHAN, T. PCA and Autoencoders. 2017.

MASCI, J. et al. **Stacked convolutional auto-encoders for hierarchical feature extraction**. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). **Anais**...Springer, Berlin, Heidelberg, 2011Disponível em: <https://link.springer.com/chapter/10.1007/978-3-642-21735-7_7>. Acesso em: 7 set. 2020

MELO, P. M. A. S. Conceitos básicos da meta-heurística Tabu Search. **Faculdade de Engenharia da Universidade do Porto**, 2008.

MISRA, S. et al. **An Autoencoder Based Model for Detecting Fraudulent Credit Card Transaction**. Procedia Computer Science. **Anais**...Elsevier B.V., 1 jan. 2020

MOHAPATRA, S.; PATRA, D.; SATPATHY, S. An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. **Neural Computing and Applications**, v. 24, n. 7, p. 1887–1904, 2014.

MURPHY, K. R. et al. Fluorescence spectroscopy and multi-way techniques. PARAFAC. **Analytical Methods**, v. 5, n. 23, p. 6557–6566, 2013.

MYSZKOWSKI, P. B. et al. Hybrid ant colony optimization in solving multi-skill resource-constrained project scheduling problem. **Soft Computing**, v. 19, n. 12, p. 3599–3619, 2015.

O. DUDA, R.; E. HART, P.; G.STORK, D. Pattern Classification / R.O. Duda, P.E. Hart, D.G. Stork. 2019.

OSTERTAGOVÁA, E. Modelling using polynomial regression. 2012.

PALUSZKIEWICZ, C. et al. **FT-IR study of montmorillonite-chitosan nanocomposite materials**. Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy. **Anais**...15 ago. 2011

PES, B. Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains. **Neural Computing and Applications**, 2019.

PESSOA, C. M. et al. Development of Ant Colony Optimization (ACO) Algorithms Based on Statistical Analysis and Hypothesis Testing for Variable Selection. **IFAC-PapersOnLine**, v. 48, n. 8, p. 900–905, 2015.

RANZAN, C. et al. Wheat flour characterization using NIR and spectral filter based on Ant Colony Optimization. **Chemometrics and Intelligent Laboratory Systems**, v. 132, n. 0, p. 133–140, 2014.

RANZAN, C. et al. Sulfur Determination in Diesel using 2D Fluorescence Spectroscopy and Linear Models. **IFAC-PapersOnLine**, v. 48, n. 8, p. 415–420, 2015.

RANZAN, L. et al. Classification of Diesel Fuel Using Two-Dimensional Fluorescence Spectroscopy. **Energy & Fuels**, v. 31, n. 9, p. 8942–8950, 2017.

REID, L. M.; O'DONNELL, C. P.; DOWNEY, G. Recent technological advances for the determination of food authenticity. **Trends in Food Science and Technology**, v. 17, n. 7, p. 344–353, 2006.

REUNANEN, J.; GUYON, I.; ELISSEEFF, A. Overfitting in Making Comparisons Between Variable Selection Methods. **Journal of Machine Learning Research**, v. 3, p. 1371–1382, 2003.

RÍOS-REINA, R. et al. Characterization and authentication of Spanish PDO wine vinegars using multidimensional fluorescence and chemometrics. **Food Chemistry**, v. 230, p. 108–116, 2017.

ROBERTSON, A. L. et al. **Using a partial least squares (PLS) method for estimating cyanobacterial pigments in eutrophic inland waters**. Remote Sensing and Modeling of Ecosystems for Sustainability VI. **Anais**...SPIE, 20 ago. 2009

ROBNIK-ŠIKONJA, M.; KONONENKO, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. **Machine Learning**, v. 53, n. 1, p. 23–69, 2003.

RODRIGUES E SILVA, S.; SCHIMIDT, F. Redução de variáveis de entrada de redes neurais artificiais a partir de dados de análise de componentes principais na modelagem de oxigênio dissolvido. **Quimica Nova**, v. 39, n. 3, p. 273–278, 1 abr. 2016.

RUTHERFORD, J. W. et al. Excitation emission matrix fluorescence spectroscopy for combustion generated particulate matter source identification. **Atmospheric Environment**, v. 220, p. 117065, 2020.

RYGULA, A. et al. **Raman spectroscopy of proteins: A reviewJournal of Raman Spectroscopy**, 2013.

SAKURADA, M.; YAIRI, T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. **Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis**, p. 4, 2014.

ŠAŠIĆ, S. **Pharmaceutical Applications of Raman Spectroscopy**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2007.

SCHERER, D.; MÜLLER, A.; BEHNKE, S. **Evaluation of Pooling Operations in Convolutional Architectures for Object Recognition**. [s.l: s.n.]. Disponível em: <http://www.ais.uni-bonn.de>. Acesso em: 7 set. 2020.

SCHWENK, H.; BENGIO, Y. Boosting Neural Networks. **Neural Computation**, v. 12, n. 8, p. 1869–1887, 2000.

SEBBEN, J. A. et al. Development of a quantitative approach using Raman spectroscopy for carotenoids determination in processed sweet potato. **Food Chemistry**, v. 245, p. 1224–1231, 2018.

SERBENCU, A.; MINZU, V. **Hybridized Ant Colony System for Tasks to Workstations Assignment**. 2016 IEEE Symposium Series on Computational Intelligence (SSCI).

**Anais**...2016

SETTLE, F. **Handbook of Instrumental Techniques for Analytical Chemistry**. Upper Saddle River: Prentice-Hall, 1997.

SHINZAWA, H. et al. Accelerated weathering-induced degradation of poly (lactic acid) fiber studied by near-infrared (NIR) hyper spectral imaging. **Applied Spectroscopy**, v. 66, n. 4, p. 470–474, 2012.

SILVA, A. C. et al. Green chemistry method based on PARAFAC EEM data modeling for Benzo[a]pyrene quantitation in distilled spirit. **Journal of the Brazilian Chemical Society**, v. 30, n. 2, p. 398–405, 2019.

SILVA, A. C. DA et al. Two-dimensional linear discriminant analysis for classification of three-way chemical data. **Analytica Chimica Acta**, v. 938, p. 53–62, 2016.

SILVA, J. I. S. DA; SECCHI, A. R. AN APPROACH TO OPTIMIZE COSTS DURING ULTRA-LOW HYDRODESULFURIZATION OF A BLEND CONSISTING OF DIFFERENT OIL STREAMS. **Brazilian Journal of Chemical Engineering**, v. 35, p. 1293–1304, 2018.

SIMONYAN, K.; ZISSERMAN, A. **VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION**. [s.l: s.n.]. Disponível em: <http://www.robots.ox.ac.uk/>. Acesso em: 24 ago. 2020.

SINELLI, N. et al. Evaluation of quality and nutraceutical content of blueberries (Vaccinium corymbosum L.) by near and mid-infrared spectroscopy. **Postharvest Biology and Technology**, v. 50, n. 1, p. 31–36, out. 2008.

SKOOG, D. A.; HOLLER, J. F.; CROUCH, S. R. **Princípios de Análise Instrumental**. 6ª edição ed. [s.l.] Bookman, 2009.

SORIANO-DISLA, J. M. et al. **The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological propertiesApplied Spectroscopy Reviews**, 17 fev. 2014.

SRINIVAS, S.; SUBRAMANYA, A.; BABU, R. V. Training Sparse Neural Networks. **IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops**, v. 2017- July, p. 455–462, 21 nov. 2016.

SRIVASTAVA, N.; MANSIMOV, E.; SALAKHUTDINOV, R. Unsupervised Learning of Video Representations using LSTMs. **32nd International Conference on Machine Learning, ICML 2015**, v. 1, p. 843–852, 16 fev. 2015.

STÜTZLE, T.; LÓPEZ-IBÁÑEZ, M.; DORIGO, M. A Concise Overview of Applications of Ant Colony Optimization. In: [s.l: s.n.].

TANG, J.; ALELYANI, S.; LIU, H. Feature selection for classification: A review. In: **Data Classification: Algorithms and Applications**. [s.l: s.n.]. p. 37–64.

TIPPING, M. E.; BISHOP, C. M. Probabilistic Principal Component Analysis. **Journal of the Royal Statistical Society. Series B (Statistical Methodology)**, v. 61, n. 3, p. 611–622, 1999.

VALENTI, B. et al. Infrared spectroscopic methods for the discrimination of cows' milk according to the feeding system, cow breed and altitude of the dairy farm. **International Dairy Journal**, v. 32, n. 1, p. 26–32, 2013.

VERMA, K.; KUMAR SINGH, P. An Insight to Soft Computing based Defect Prediction Techniques in Software. **International Journal of Modern Education and Computer Science**, v. 7, n. 9, p. 52–58, 8 set. 2015.

VINCENT, P. et al. Extracting and composing robust features with denoising autoencoders. **Proceedings of the 25th international conference on Machine learning - ICML '08**, p. 1096–1103, 2008.

WANG, Y. et al. An Improved Ant Colony Optimization algorithm to the Periodic Vehicle

Routing Problem with Time Window and Service Choice. **Swarm and Evolutionary Computation**, v. 55, p. 100675, 1 jun. 2020.

WETZEL, S. J. Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. **Physical Review E**, v. 96, n. 2, p. 22140, 2017.

WIDROW, B.; LEHR, M. A. 30 Years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation. **Proceedings of the IEEE**, v. 78, n. 9, p. 1415–1442, 1990.

WILSON, J. H.; ZHANG, C.; KOVACS, J. M. Separating crop species in Northeastern Ontario using hyperspectral data. **Remote Sensing**, v. 6, n. 2, p. 925–945, 2014.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 58, n. 2, p. 109–130, 2001.

XIAOBO, Z. et al. Variables selection methods in near-infrared spectroscopy. **Analytica Chimica Acta**, v. 667, n. 1, p. 14–32, 2010.

XU, L. et al. Variable-weighted PLS. **Chemometrics and Intelligent Laboratory Systems**, v. 85, n. 1, p. 140–143, 2007.

YADAV, L. D. S.; YADAV, L. D. S. Infrared (IR) Spectroscopy. In: **Organic Spectroscopy**. [s.l.] Springer Netherlands, 2005. p. 52–106.

YAMASHITA, R. et al. **Convolutional neural networks: an overview and application in radiologyInsights into Imaging**Springer Verlag, , 1 ago. 2018. Disponível em: <https://link.springer.com/articles/10.1007/s13244-018-0639-9>. Acesso em: 7 set. 2020

YU, C.; YAO, W. Robust linear regression: A review and comparison. **Communications in Statistics - Simulation and Computation**, v. 46, n. 8, p. 6261–6282, 14 set. 2017.

YU, D. et al. Near-infrared fluorescent probe for detection of thiophenols in water samples and living cells. **Analytical Chemistry**, v. 86, n. 17, p. 8835–8841, 2014.

ZEILER, M. D.; FERGUS, R. **Visualizing and Understanding Convolutional Networks**. (D. Fleet et al., Eds.)Computer Vision – ECCV 2014. **Anais**...Cham: Springer International Publishing, 2014

ZHANG, G.; LI, H. **Effectiveness of Scaled Exponentially-Regularized Linear Units (SERLUs)**. [s.l: s.n.].

ZHANG, P. (ED.). Front Matter. In: **Advanced Industrial Control Technology**. Oxford: William Andrew Publishing, 2010. p. iii.

ZHANG, X.; TANG, L. **A New Hybrid Ant Colony Optimization Algorithm for the Traveling Salesman Problem**. (D.-S. Huang et al., Eds.)Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2008

ZHENG, C. et al. Feature selection for high-dimensional integrated data. In: **Proceedings of the 2012 SIAM International Conference on Data Mining**. Proceedings. [s.l.] Society for Industrial and Applied Mathematics, 2012. p. 1141–1150.

ZOU, J.; ZHANG, J.; JIANG, P. Credit Card Fraud Detection Using Autoencoder Neural Network. 30 ago. 2019.

# Capítulo 3 – **Prediction of Sulfur Content in Diesel Fuel using Fluorescence Spectroscopy and a Hybrid Ant Colony - Tabu Search Algorithm with Polynomial Basis Expansion**

**Abstract**[1]: It is widely accepted that feature selection is an essential step in predictive modeling. There are several approaches to feature selection, from filter techniques to meta-heuristics wrapper methods. In this paper, we propose a compilation of tools to optimize the fitting of black-box linear models. The proposed AnTSbe algorithm combines Ant Colony Optimization and Tabu Search memory list for the selection of features and uses l1 and l2 regularization norms to fit the linear models. In addition, a polynomial combination of input features was introduced to further explore the information contained in the original data. As a case study, excitation-emission matrix fluorescence data were used as the primary measurements to predict total sulfur concentration in diesel fuel samples. The sample dataset was divided into S10 (less than 10 ppm of total sulfur), and S100 (mean sulfur content of 100 ppm) groups and local linear models were fit with AnTSbe. For the Diesel S100 local models, using only 5 out of the original 1467 fluorescence pairs, combined with basis expansion, we were able to satisfactorily predict total sulfur content in samples with MAPE of less than 4% and RMSE of 4.68 ppm, for the test subset. For the Diesel S10 local models, the use of 4 Ex/Em pairs was sufficient to predict sulfur content with MAPE 0.24%, and RMSE of 0.015 ppm, for the test subset. Our experimental results demonstrate that the proposed methodology was able to satisfactorily optimize the fitting of linear models to predict sulfur content in diesel fuel samples without need of chemical of physical pre-treatment, and was superior to classic PLS regression methods and also to our previous results with ant colony optimization studies in the same dataset. The proposed AnTSbe can be directly applied to data from other sources without need for adaptations.

[1] Capítulo referente ao artigo publicado no periódico *Chemometrics and Intelligent Laboratory Systems*, setembro 2020. https://doi.org/10.1016/j.chemolab.2020.104161

## 3.1 Introduction

Nowadays, predictive modeling permeates every knowledge field. The ability to quantify an output of interest, classify occurrences, or predict future outcomes using auxiliary measures are base principles of many technological advances of the last decades. From detecting cancer (EINIPOUR; CORRESPONDING, 2011; SUN et al., 2019), social media behavior (MORO; RITA; VALA, 2016), to industrial production (BELTRAMO; KLOCKE; HITZMANN, 2019), new machine learning algorithms are changed and developed daily to predict information that would be too costly, invasive, or impossible to access directly.

Black box models establish a functional relationship between system inputs and outputs (GUIDOTTI et al., 2018). The parameters of these functions do not need to have any phenomenological significance (e.g., heat or mass transfer coefficients or reaction kinetics), but are very efficient in faithfully representing trends in the process behavior (ZHANG, 2010). With a collection of empirical or simulated data of a system, a model can be fitted to find the correlation of that information with one or more outputs of interest.

In many applications, the number of available input features can reach hundreds or even thousands of variables (e.g., image classification (JOHNSON; XIE, 2013; KAVZOGLU, 2017; MOHAPATRA; PATRA; SATPATHY, 2014), spectral data (RANZAN et al., 2014, 2017; SEBBEN et al., 2018), and industrial processes (SILVA; SECCHI, 2018; ZHENG et al., 2012)). However, data can be associated with a high level of noise, collinearity, and be filled with irrelevant or redundant variables (TANG; ALELYANI; LIU, 2014). Several selection techniques were developed to address the problem of extracting valid information from the features which can efficiently describe the input data while reducing noise and useless variables (CHANDRASHEKAR; SAHIN, 2014). These techniques can be categorized into Feature extraction and Feature selection.

Feature extraction approaches project features into a new space with lower dimensionality by combining the original feature space. Examples include Principal Component Analysis (PCA) (CAMACHO; PICÓ; FERRER, 2010; TIPPING; BISHOP, 1999), Linear Discriminant Analysis (LDA) (SILVA et al., 2016), and Partial Least Squares (PLS) (WOLD; SJÖSTRÖM; ERIKSSON, 2001), which are the most widely applied techniques to deal with high dimensionality data (GODOY; VEGA; MARCHETTI, 2014). However, it is difficult to link the features from the original space with the new features since there is no physical meaning for the transformed variables.

On the other hand, the Feature selection approaches aim to select a small subset of features that minimize redundancy and maximize relevance to the target, maintaining the physical meanings (TANG; ALELYANI; LIU, 2014). Metaheuristics search algorithms (as Genetic Algorithm (GOLDBERG, 1989), Particle Swarm Optimization (HU et al., 2019) and Ant Colony Optimization (PESSOA et al., 2015)) use the predictor as a black box and the predictor performance as the objective function to evaluate a feature subset. The feature search component will produce a set of features that will be used by the learning algorithm to predict an output. The performance of the prediction returns to the feature search component for the next iteration of subset selection.

The Ant Colony Optimization (ACO) is a metaheuristic multi-agent algorithm used to solve hard combinatorial optimization problems. There are several review and survey

papers (CHANDRA MOHAN; BASKARAN, 2012; DORIGO; BLUM, 2005; STÜTZLE; LÓPEZ-IBÁÑEZ; DORIGO, 2011) dedicated to ACO applications, that include areas as fluid dynamics, telecommunications, bioinformatics, system modeling, simulation, image processing, routing, scheduling, and production problems, logistics, transportation and supply chain management. ACO algorithms try to mimic the food foraging behavior of real ants adapting a pheromone memory model that governs the way agents wander the search space. The management of this pheromone influences the diversification (i.e., exploration) and the intensification (i.e., exploitation) of the search process. A well-design ACO algorithm uses strategies to balance exploration and exploitation (E&E) to find high-quality solutions for problems (BULLNHEIMER; HARTL; STRAUSS, 1999; GAMBARDELLA; DORIGO, 2000; KU-MAHAMUD; ALOBAEDY, 2013; RANZAN et al., 2014). A natural evolution of E&E techniques was the hybridization of metaheuristic algorithms (BLUM et al., 2008, 2011). The incorporation of principles from other searching algorithms into ACO was used to solve the Traveling Salesman Problem (ZHANG; TANG, 2008), Vehicle Routing problem (ABDULKADER; GAJPAL; ELMEKKAWY, 2015; BALSEIRO; LOISEAU; RAMONET, 2011; LI; TIAN; LEUNG, 2009), Tasks to Workstations Assignment problem (SERBENCU; MINZU, 2016), Multi-skill Resource-constrained Project Scheduling problem (MYSZKOWSKI et al., 2015), and Quadratic Assignment problems (ARITO; LEGUIZAMÓN, 2009; TSUTSUI; FUJIMOTO, 2011).

The two major contributions of our methodology is the proposal of a hybrid variable selection algorithm based on Ant Colony Optimization and Tabu Search (TS) (NIU et al., 2018; PIRIM; BAYRAKTAR; EKSIOGLU, 2008), to solve early stagnation and avoid redundant calculations, and the use of the expansion of bases to further explore the information contained in the data. After the selection of inputs (and before model fitting), the selected variables are expanded as a new feature matrix consisting of all polynomial combinations of features with degree equal or less than a defined value. This expansion can capture non-linear and combinatorial information that may not be perceived otherwise.

To evaluate the methodology, a case study is presented, where Excitation-Emission Matrix (EEM) fluorescence spectroscopy is used as primary information to predict total sulfur concentration in diesel fuel.

## 3.2 Methodology

All implementations in this work were done in Python v3.5.4.1 in combination with the readily available modules, especially from the SciKit Learn library version 0.20.2 54.

### *3.2.1 Preprocessing*

The first necessary step for any modeling procedure is the pre-treatment of the data. If misleading information, outliers, and unscaled data are given to the optimizer, there will be a detriment in the efficiency of the algorithm. The preprocessing routine comprehends **i.** outliers detection; **ii.** data segmentation into training, validation, and test data sets; and **iii.** data scaling.

#### *3.2.1.1 Outliers Detection – Hoteling's $T^2$ Statistic*

This method is coupled with Principal Component Analysis (PCA). All input data is first scaled to mean zero and standard deviation one; then, PCA is applied, and the eigenvalues

and eigenvectors are used to calculate the $T^2$ statistic for each sample, expressed as Eq. 3.1 (RUSSELL; CHIANG; BRAATZ, 2000):

$$T^2 = X^T \widehat{W} \hat{\Lambda}^{-1} \widehat{W}^T X \tag{3.1}$$

where $\hat{\Lambda}$ = diag($\lambda_1$, $\lambda_2$, ..., $\lambda_l$) is a diagonal matrix containing the eigenvalues related to the *l* retained PCs; X the matrix of scaled inputs; and $\widehat{W}$ the matrix of the *l* retained eigenvectors. In this implementation, the number of retained PCs is the one able to capture at least 95% of the original variance of the data. The $T_\alpha^2$ threshold is computed as Eq. 3.2:

$$T_\alpha^2 = \frac{l(N-1)}{(N-l)} F_{l,N-l,\alpha} \tag{3.2}$$

where *l* is the number of retained principal components, *N* the number of samples, α the level of significance (defined as 5%) and $F_{l,N-l,\alpha}$ is the Fisher distribution with *l* and (N-*l*) degrees of freedom. Any sample with $T^2$ higher than $T_\alpha^2$ is considered an outlier and removed from the dataset (MANSOURI et al., 2016).

#### *3.2.1.2 Dataset splitting based on a modified version of K-rank*

The methodology for splitting the data into calibration (cal), validation (val), and testing (test) subsets is the one implemented by Santos *et al.* (SANTOS et al., 2019). This methodology is especially useful when dealing with multiple solutions problems: situations where multiple combinations of the input variables can yield the same output y.

First, the user chooses a k number of clusters, ranging from 1 to N-1 (number of samples - 1). Then, a k-means algorithm (RASCHKA, 2015) with k centroids is run using only the input variables to split the dataset into ki similar groups. For each cluster: the $k_{i,samples}$ are sorted in ascending order for a selected output y; the proportions of each subset are chosen (e.g., 60% calibration – 20% validation – 20% testing), and the methodology adapts a pattern to select, in order, the samples to their respective subsets (e.g., cal-cal-cal-val-test …). In this implementation, the extremes samples (with minimum and maximum values of y) in each cluster are always selected for the training subset, to avoid extrapolation. This clustering of data is especially useful when dealing with cases that have a multiplicity of solutions.

Although other splitting methodologies (as cross-validation) can be appealing when fitting black-box models (to avoid overfitting and the influence of abnormal samples), the selection of variables can be a very time-demanding task, and the use of multiple cal/val/test subsets can, for some cases, render the total computational time prohibitive.

#### *3.2.1.3 Data Scaling*

Standardization is a common requirement for many machine learning estimators. Many elements used in the objective function of a learning algorithm (as *l*1 and *l*2

regularizers of linear models) assume that all features are centered on zero and have variance in the same order. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn correctly from other features as expected (PEDREGOSA et al., 2011). The scaler function must be fitted using the training subset, and then the other subsets are scaled accordingly. The user can define which scaler is better for their specific case.

After the pre-processing stage, we end up with clean and scaled data, with all samples already divided into subgroups, ready to be processed by the AnTSbe algorithm, the core optimizer that will venture through the possible combinations of input features to fit linear models and predict the desired output.

### *3.2.2 AnTSbe – Ant Colony Optimizer hybridized with Tabu Search and basis expansion*

The AnTSbe algorithm is based in Ant Colony Optimization (DORIGO; GAMBARDELLA, 1997), a type of stochastic optimization where multiple parallel processes (ants) take different 'routes' to minimize an objective function. When applied to optimization, the algorithm ascribes a quality indicator - called pheromone - to each input variable. The pheromone of each input is incremented after each iteration based on how well a model using this variable could predict the desired output (the lower the error, the higher the increment in pheromone). The more pheromone a variable has, the higher the chance ants in the next iteration will select it. One of the problems of swarm intelligence algorithms is the stagnation in local minimums: after a number of iterations, some inputs can dominate the pheromone trail in such a way that it loses its exploration capabilities. There are several proposals of ACO adaptations to avoid stagnation, as pheromone reset, reactive memory, smoothing, and max-min bounds (SAGBAN; KU-MAHAMUD; ABU BAKAR, 2014). In this work, we will incorporate principles from Tabu Search into the algorithm to avoid early stagnation. TS is a neighborhood search-based method that uses a memory structure to avoid being trapped in local optima. It improves the efficiency of the searching process by storing a tabu list of local solutions that were used to restrict the search by forbidding moves to some poor neighbor solutions that already have been visited (KLUABWANG; PUANGDOWNREONG; SUJITJORN, 2012). One feature of tabu exploration is diversification, responsible for moving the exploration process over different regions of the search space. In this implementation, TS is hybridized with ACO not to search for neighbor solutions directly, but to construct a short-term memory tabu list to avoid previously tested input combinations. The memory is short-term because it only considers tested input combinations of the last *z* iterations (after *z* iterations the input combinations are not forbidden anymore). Forbidding previously tested combinations for some iterations encourages exploration, and the evaporation of pheromone can help to avoid stagnation in local optima.

To compare and evaluate models, three metrics are used throughout the algorithm:

(i) Mean Absolute Percentage Error (MAPE), defined as Eq. 3.3:

$$MAPE = \frac{100\%}{n}\sum_{i}^{n}\left|\frac{(y_i - \hat{y}_i)}{y_i}\right| \tag{3.3}$$

where *n* is the number of samples, *y* the real value of the output and $\hat{y}$ the predicted value of the output;

(ii) Root Mean square error (RMSE) – Eq. 3.4,

$$RMSE = \sqrt{\frac{1}{n}\sum_{i}(y_i - \hat{y}_i)^2} \tag{3.4}$$

and (iii) the Coefficient of Determination ($R^2$) – Eq. 3.5,

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{3.5}$$

where $\bar{y}$ is the mean value of y.

The AnTSbe is divided into three phases: Phase One is the initialization stage, where the optimization parameters are chosen; Phase Two is the core of the optimization, where models are fitted, evaluated and compared; and, in Phase Three, the best-fitted models are presented, with their corresponding variables and all comparative metrics.

#### *3.2.2.1 Phase One – Initialization*

The initialization process starts with the selection of all needed optimization parameters:

**Model Size** – *Nw*: the number of input variables in each model. The user can define $Nw_0$ and $Nw_n$ to be the initial and final model sizes, respectively (with $Nw_0 \leq Nw_n$). If defined, *Nw* will assume values within $Nw_0$ and $Nw_n$. The algorithm as a whole (Phase one to three) will be run individually for each *Nw*.

**Type of model**: the type of regression to be fitted. The Ridge Regression (through the scikit-learn function sklearn.linear_model.Ridge) solves a regression model where the loss function is the Linear Least Squares function, and the regularization is given by the *l*2-norm, which aims to reduce the magnitude of coefficients (TIKHONOV; ARSENIN, 1977). The Lasso Regression, similarly, uses *l*1-norm regularization, which penalizes the number of total parameters in the model, reducing some coefficients to absolute zero (MASSARON; BOSCHETTI, 2016). The used Scikit-learn LassoLarsIC (criterion='bic') implementation solves the Lasso model using Least Angle Reduction (Lars), and the selection of the regularization parameter α is based on the Bayesian information criterion (EFRON et al., 2004), making a trade-off between the goodness of fit and the complexity of the model. The use of regularization will foment the fitting of models more robust to overfitting, discarding inefficient input variables and avoiding singularity issues during a model fitting in cases where there are more variables than observations.

**Base Expansion – σ**: generate a new feature matrix consisting of all polynomial combinations of features with degree equal or less than σ. Even though the models are linear in the parameters, the expansion of the selected input variables into polynomial combinations can capture information that could improve the prediction metrics. It was essential to only expand the bases after the selection of inputs, because, in cases where there are thousands of variables, if we simply expand all the features before selection, the complexity of the problem and the number of local minimums would increase exponentially.

**Optimization Metric – *OptMetric***: the metric involved in the loss function the algorithm will minimize. It can be RMSE or MAPE. If MAPE is selected, it has the tendency to minimize prediction errors of outputs with values closer to zero. If RMSE is selected, it has the tendency to minimize prediction errors of outputs with higher values.

**Number of runs – μ**: number of times the algorithm resets the pheromone trail to its initial value $\tau_0$.

**Number of iterations – *t***: total number of iterations the algorithm performs in each run.

**Number of ants – $N_{ants}$**: number of models fitted in each iteration.

**Tabu memory size – *z***: number of past iterations where the tested combinations are part of the tabu memory.

**Initial pheromone value – $\tau_0$**: the initial amount of pheromone for each variable.

**Pheromone gain – *k***: numerator of the expression of pheromone increment each ant will add to the variables it has selected in that iteration.

**Pheromone evaporation rate – ρ**: defines how much of the current pheromone will not be kept for the next iteration. If close to one, it will heavily penalize unselected inputs or inputs that fitted models with high prediction errors.

Once all parameters are established, the Global Solution is initialized by fitting a model with random *Nw* variables. This Global Solution will be latter compared to future fitted models. The tabu memory is also initialized as an empty list.

#### *3.2.2.2 Phase Two – Optimization*

During Phase Two, the ant army will evaluate possible combinations of input variables that could minimize the loss function. At the beginning of each $t_i$ iteration, each of the $N_{ants}$ ants will select *Nw* input variables. This selection is based on the pheromone trail and in a random factor. The random factor is represented by a random trigger, generating values between 0 and 1. The pheromone trail is transformed in a pheromone density vector *γ*, by dividing the pheromone of each input by the sum of all pheromone, and then accumulated as C*γ* (Eq. 3.6):

$$C\gamma_i = \sum_{1}^{i} \gamma_i \quad \therefore \quad \gamma_i = \frac{\tau_i}{\sum \tau} \tag{3.6}$$

The random trigger is fired, and its value compared to the accumulated pheromone density. The closest input with accumulated density higher than the trigger is selected, removed from the selection pool, and the *γ* is updated without that input. The procedure is repeated until *Nw* features are selected. The combination between the random trigger and the density of pheromone guarantees that every input has a chance into being selected, but the higher the pheromone density, the greater the chance of an input being selected. Figure 3.1 shows an example of input selection with a random trigger of 0.856. In this example, input $X_9$ would be the one selected.

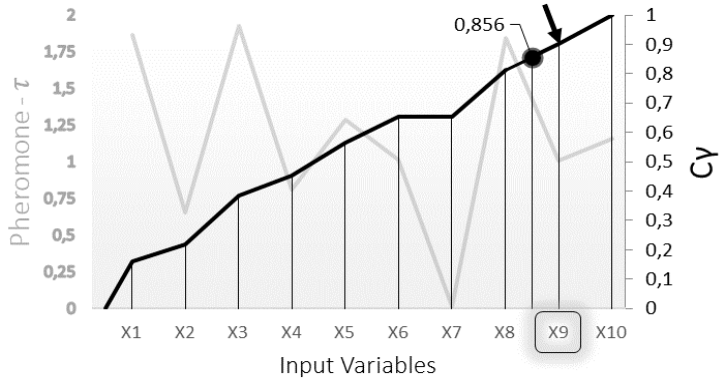



Figure 3.1. Example of pheromone-based input selection with a random trigger of 0.856.

After the ant has selected *Nw* inputs, its selection is compared to the combinations stored in the tabu memory. If the ant has selected a forbidden combination, the ant resets and starts the selection once again. If the ant combination is not forbidden, the selected inputs are expanded as all polynomial combination of features with degree equal or less than σ (*e.g.*, if [$X_1$ , $X_2$] are the chosen inputs, an expansion with σ = 2 will result in a new matrix [$X_1$, $X_2$, $X_1^2$, $X_2^2$, $X_1$*$X_2$]).

The ant will use the chosen variables (original or expanded, depending on the σ) and samples from the calibration subgroup to fit a linear model of the selected regression type. The fitted model is used to predict the output of interest for all subgroups (calibration, validation, and test), and all metrics are evaluated. The ant compares the quality of its model with the Global Solution, based on a loss function $\psi_q$ (Eq. 3.7),

$$\begin{aligned} \psi_q &= [OptMetric]_{\text{cal}} + [OptMetric]_{\text{val}} + \Gamma \\ \Gamma &= \frac{\max\langle [OptMetric]_{cal}, [OptMetric]_{val} \rangle + 1}{\min\langle [OptMetric]_{cal}, [OptMetric]_{val} \rangle + 1} \end{aligned} \tag{3.7}$$

if the ant's $\psi_q$ is smaller than the Global Solution's $\psi_q$, the ant's model becomes the new Global Solution.

Finally, the ant deposits pheromone in each of the variables it has selected according to Eq. 3.8:

$$\tau_{i,t+1} = \tau_{i,t} + \frac{k}{(\psi_q + 1)^3} \tag{3.8}$$

where *k* is the pheromone gain. The higher the prediction error the model has, the smaller the increment in pheromone. If by any chance (as in regressions with *l1-norm*), the fitted parameter of an input is absolute zero, then no pheromone is added to that variable. As all ants run in parallel, the increment of pheromone each ant deposits in their selected variables will only be perceived in the next iteration, not affecting the variable selection of other ants in the current $t_i$.

After all $N_{ants}$ have fitted their models and deposited their pheromones, the pheromone trail is evaporated, multiplying the trail by the evaporation rate ρ (Eq. 3.9):

$$\tau_{t+1} = \tau_{t+1} \cdot (1 - \rho) \tag{3.9}$$

The tabu memory list is updated, adding all $N_{ants}$ combinations tested in the current iteration (and removing combinations from any other than the last *z* iterations).

The routine is repeated *t* times, and at each iteration, the pheromone trail is updated. The final pheromone trail can indicate which of the input variables had a higher correlation with the desired output, been part of models with smaller predictive errors. To avoid local minima, Phase Two is repeated μ times, each time restarting the pheromone trail to the initial $\tau_0$ (reset pheromone memory) and emptying the tabu memory list, but carrying on the Global Solution. At the end of every μ run, the final pheromone trail – $\tau_{F\mu}$ – and the best predictive model of each particular run are saved for future reference.

#### *3.2.2.3 Phase Three – Global Solution*

After μ runs, each with *t* iterations and $N_{ants}$ models fitted at each iteration, the algorithm returns the Global Solution - the fitted model with lowest $\psi_q$ - along with its metrics for all subgroups, selected input variables, and corresponding parameters. If the metrics of the test subset are equivalent to the calibration and validation metrics, then the fitted model can be considered robust when dealing with samples never seen before. In addition, the global pheromone trail – $\tau_G$ – is presented, as the sum of all $\tau_F$, normalized between 0 – 1 (dividing the vector by its maximum value), as can be seen in Eq. 3.10. The global pheromone trail is an indicator of which input variables had greater success predicting the output of interest throughout the optimization, been part of models that had better metrics and smaller prediction errors.

$$\tau_G = \frac{\sum_1^{\mu} \tau_F}{\max\left(\sum_1^{\mu} \tau_F\right)} \tag{3.10}$$

Depicted in Figure 3.2 we can see a schematic representation of the AnTSbe algorithm.

If different model sizes were fitted, the user could select the one that fits better to their needs, based on all calculated metrics for each of the returned Global Solutions. Ideally, if there is a clear correlation between the inputs and the desired output, different sized models should mostly select the same variables.

In cases where there are many highly correlated input variables, the AnTSbe algorithm can be used as a variable filter for further analyses. After running the AnTSbe, it is possible to filter the input variables that had greater success in predicting the desired output by selecting the inputs with higher pheromone (quality indicator) concentration. If the user wants to filter inputs from optimization with only one specific set of parameters, a filter vector is created directly as the global pheromone trail of that optimization. If the user wants to filter inputs considering multiple sets of parameters (*e.g.,* various model sizes for *Ridge* models and the same σ), a multi-filter vector is created as the sum of all individual global pheromone trails. The filter/multi-filter vector is sorted in descending order, and the user defines how many of the first variables should be selected as filtered inputs.

With these filtered inputs, the AnTSbe can be re-run, with any selected parameters, but with the restriction of only selecting inputs within this filtered base. This procedure can be recursively done.

To evaluate the proposed methodology, a case study will be presented next, making use of excitation−emission matrix fluorescence spectroscopy to quantify sulfur concentration in diesel fuel.

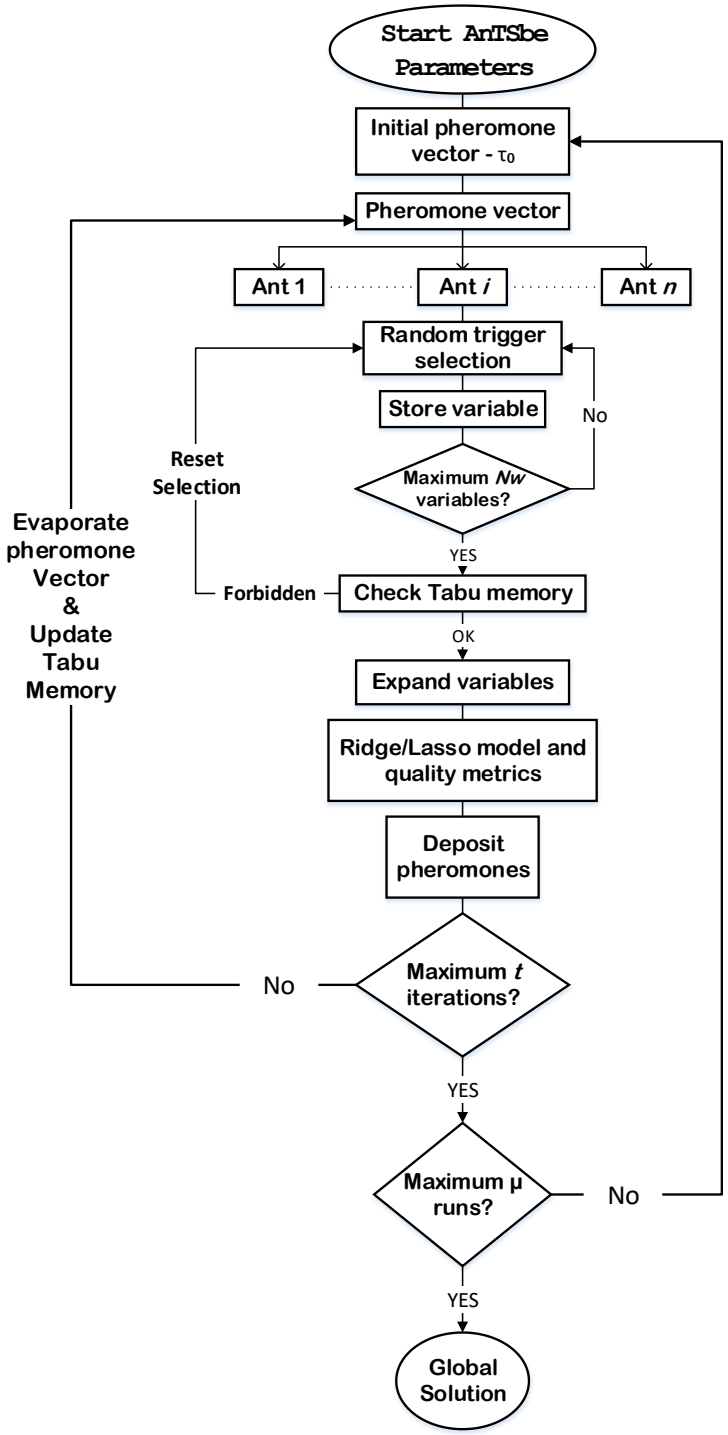



Figure 3.2. Schematic representation of the AnTSbe algorithm.

## 3.3 Case Study – Quantifying total sulfur content in diesel fuel samples using EEM fluorescence spectroscopy

The dispersion of sulfur oxides in the atmosphere from the combustion of fossil fuels is a sensitive topic in environmental laws (CHANDRA SRIVASTAVA, 2012). The sulfur contained in the fuel is directly responsible for the emission of sulfate particulates and $SO_3$/$SO_2$ (that causes acid rain) during combustion, the poisoning of refining catalysts, and the corrosion of pipes, storage units and motors (MURAKAMI, 1995). In the last 20 years, most developed countries severely changed their legislations regarding maximum sulfur content in fuels, going from ten thousand ppm to near-zero levels, being less than 15 ppm the typical limit for transportation diesel and gasoline worldwide (LIMA et al., 2018).

Hydrodesulphurization (HDS) is the most common catalytic chemical process used to remove sulfur from refined petroleum products. In a refinery, the reaction takes place in a fixed-bed reactor at elevated temperatures (from 300 to 400°C) and elevated pressures (30 – 130 atmospheres of absolute pressure), typically using alumina base impregnated with cobalt and molybdenum as a catalyst, combining the feed with a hydrogen-rich stream and retrieving hydrogen sulfide ($H_2S$) (BRUNET et al., 2005). HDS is a costly procedure and should be fine-tuned to avoid inefficiency, producing streams out of specification that demand rework.

Presently, to quantify total sulfur concentration in a stream after HDS, a sample must be collected and taken to a laboratory where qualified people will handle it and run the specific certification test for sulfur quantification in their legislation (as the ASTM D-4294 (ASTM, 2010), in Brazil). This quantification procedure, besides being expensive, usually takes hours in a typical refinery. That means that when the operator finally receives feedback, it can be too late to make any control action.

To implement advanced controlling techniques to the HDS, it is of utmost importance to be able to predict in real-time the sulfur concentration of the desulfurized product. As can be seen in works as (ABURTO et al., 2014; CAMPOS et al., 2018; RANZAN et al., 2015, 2017), EEM fluorescence spectroscopy can be used as a non-invasive, fast, and sensitive technique to capture information about total sulfur content in diesel fuel. After HDS, most of the oil's sulfur is contained in stable polycyclic aromatic molecules as benzothiophenes and dibenzothiophenes, being the latter the hardest sulfur to remove in diesel oil (HUA et al., 2003; WANG et al., 2003). These compounds and their derivatives present luminous proprieties, being natural fluorophores (AARON et al., 2002; BREE; ZWARICH, 1971; HOU et al., 2019; NAYAK; AGARWAL; PERIASAMY, 2010)

The objective of this case study is to apply the AnTSbe methodology to optimize the selection, among thousands of fluorescence excitation-emission pairs, of the ones able to predict total sulfur content in diesel samples satisfactorily.

### *3.3.1 Dataset*

The dataset used in this study is the one gathered by Ranzan *et al.* (2017): excitation-emission matrix fluorescence spectra of sixty-one samples of diesel fuel, provided and certified by a Brazilian petroleum refinery. The samples were characterized as Diesel S10 (samples with total sulfur lower than 10 ppm) and Diesel S100 (samples with average total sulfur around 100 ppm). The referred work used the data in a purely classificatory study, being able to label all samples correctly, but without any regard about sulfur quantification.

**Diesel S10** – eleven samples were characterized as Diesel S10; they had between 5.1 and 6.4 ppm of total sulfur, with an average of 5.8 ppm. The sulfur content of these samples was certified according to the ASTM Standard D-7039 (ASTM, 2015), using a Sindie 7039 bench analyzer by XOS®.

**Diesel S100** – fifty samples were characterized as Diesel S100, with total sulfur content between 73.7 and 118 ppm, and an average of 99.5 ppm. All samples were certified according to the ASTM Standard D-4294 (ASTM, 2010), using a LABX-3000 by Oxford®.

**EEM fluorescence spectra** – the fluorescence spectra were collected using a Horiba® Fluoromax-4, equipped with a xenon lamp of 150 W. The measurements were made in a range of excitation wavelengths between 260 and 600 nm and emission wavelengths between 290 and 850 nm. The geometry of measurements was 90°. Both excitation and emission wavelengths used an increment of 10 nm. With these arrangements, each fluorescence spectra was obtained as a 57 × 35 matrix, containing the fluorescence intensity of 1995 excitation/emission (Ex/Em) pairs. As no excitation can lead to emission with a smaller wavelength, there were 1467 valid fluorescence pairs in each spectrum. Each EEM spectra was later unfolded into a row vector, the row representing the sample, and each column representing one of the Ex/Em pairs. Measurements were made in triplicate, and all samples were stabilized at 25 °C using a thermostatic bath. Figure 3.3 presents the average EEM fluorescence spectra for the Diesel S10 and Diesel S100 sample groups.

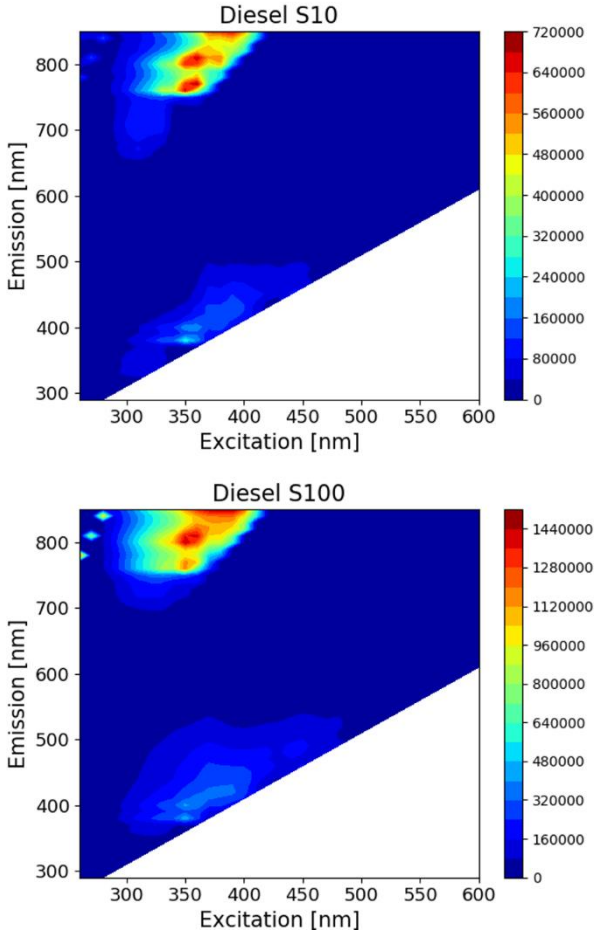



Figure 3.3. Average EEM fluorescence spectra for the Diesel S10 and Diesel S100 sample groups.

In this work, we focused on the development of local models, one for Diesel S10 and another for Diesel S100, mainly because the difference in total sulfur concentration and quantity of samples between groups was very significant. If only one global model was fitted, the Diesel S100 samples would dominate the optimization.

### *3.3.2 AnTSbe Pre-processing and Parameters*

The following parameters were applied to the optimization of both Diesel S10 and Diesel S100 local models.

For the data pre-processing, as the number of inputs vastly surpluses the number of samples, no outlier was removed in the pre-treatment. For the splitting of subgroups, total sulfur concentration (in ppm) was the selected output to sort the data, and the chosen proportions for the calibration, validation, and testing subsets were 60%, 20%, and 20%, respectively. The number of clusters was defined as one because there was no multiplicity of solutions. The *StandardScaler* was defined as the scaler function, removing the mean and scaling to unit variance (independently in each feature).

The general AnTSbe optimization parameters can be seen in Table 3.1.

Table 3.1. AnTSbe general optimization parameters for diesel fuel local models.

| **Parameter** | **Value** |
|---|---|
| OptMetric | RMSE |
| N. Runs - μ | 50 |
| N. iterations - t | 150 |
| $N_{ants}$ | 200 |
| Tabu Memory Size - z | 5 |
| Initial pheromone value – $\tau 0$ | 1000 |
| Pheromone gain – k | 100 |
| Pheromone evaporation rate – $\rho$ | 0.1 |

For the Diesel S100 models, Model Size (*Nw*) will range from 3 to 5, and the Type of Model and basis expansion (σ) were arranged as *Ridge* and σ = 1 and *LassoLarsIC* and σ = 2. With this arrangement, the algorithm as a whole will be run six times: three model sizes with *Ridge* and σ = 1 and three model sizes with *LassoLarsIC* and σ = 2. We combined the use of Lasso Regression when applying basis expansion (σ > 1), because the expansion considerably increases the number of model parameters to be fitted, and the use of *l1*-norm regularization helped to keep the models more comprehensible.

As the information contained in the EEM can be highly correlated (many Ex/Em pairs in a region could contain similar information about the desired output), this first round of experiments was also used as a filter to reduce the number of input variables. Then, the algorithm is re-run, to evaluate if we could improve the quality of the predictive models.

To filter the inputs, the multi-filter approach will be applied: the global solution for all model sizes of each specific model type will be evaluated. A multi-filter vector will be created as the sum of the global pheromone trails of each model size (3 to 5). Considering that each EEM had more than 1400 pairs, we will filter the 100 input variables with higher pheromone concentration and re-run the AnTSbe using only these filtered inputs.

For the Diesel S10 models, Model Size (*Nw*) will range from 2 to 5, and the Type of Model and basis expansion (σ) were selected as *Ridge* and σ = 1, considering the smaller number of samples. As before, the multi-filter approach will be applied to filter the 100 input variables with higher pheromone concentration and re-run the algorithm.

## 3.4 Results and Discussions

### *3.4.1 Diesel S100*

The metrics for the first optimization of the Diesel S100 local models can be seen in Table 3.2. Both the *Ridge* as the *LassoLars* models achieved similar results for the calibration and validation subsets, with MAPE smaller than 4% and RMSE around 4 ppm. For the test subset, the *Ridge* models achieved slightly better results. The base expansion increases the number of variables in the model, and, directly, the number of parameters to be fitted. Using σ = 2, the model sizes 3, 4, and 5 expand to 9, 14, and 20 variables, respectively. All *LassoLars* global solutions kept (non-zero parameters) a combination of primary and expanded variables, indicating that the use of base expansion could be beneficial for Diesel S100 predictive models. In this run, the number of non-zero parameters in *LassoLars* global solutions was 5, 6, and 7 (model sizes 3, 4, and 5).

Table 3.2. Global solutions' metrics and selected fluorescence pairs for Diesel S100 - *first optimization*.

| Diesel S100 *Ridge* and σ = 1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Calibration | | | Validation | | | Test | | |
| | | $R^2$ | MAPE | RMSE | $R^2$ | MAPE | RMSE | $R^2$ | MAPE | RMSE |
| Model Size | 3 | 0.678 | 5.16 | 6.38 | 0.876 | 2.62 | 3.50 | 0.720 | 4.25 | 4.98 |
| | 4 | 0.781 | 4.59 | 5.27 | 0.858 | 3.34 | 3.75 | 0.631 | 4.98 | 5.72 |
| | 5 | 0.823 | 3.94 | 4.73 | 0.855 | 3.14 | 3.79 | 0.758 | 3.88 | 4.64 |
| Selected Excitation/Emission Pairs | | | | | | | | | | |
| Model Size | 3 | Ex310/Em330 Ex480/Em590 Ex560/Em580 | | | | | | | | |
| | 4 | Ex290/Em350 Ex310/Em330 Ex340/Em730 Ex370/Em720 | | | | | | | | |
| | 5 | Ex310/Em330 Ex310/Em680 Ex360/Em370 Ex480/Em590 Ex560/Em580 | | | | | | | | |
| Diesel S100 *LassoLars* and σ = 2 | | | | | | | | | | |
| | | Calibration | | | Validation | | | Test | | |
| | | $R^2$ | MAPE | RMSE | $R^2$ | MAPE | RMSE | $R^2$ | MAPE | RMSE |
| Model Size | 3 | 0.857 | 3.83 | 4.26 | 0.810 | 3.39 | 4.34 | 0.480 | 5.33 | 6.80 |
| | 4 | 0.834 | 3.74 | 4.59 | 0.739 | 3.91 | 5.09 | 0.676 | 4.81 | 5.37 |
| | 5 | 0.858 | 3.85 | 4.24 | 0.882 | 2.79 | 3.42 | 0.519 | 5.39 | 6.53 |
| Selected Excitation/Emission Pairs | | | | | | | | | | |
| Model Size | 3 | Ex310/Em330 Ex370/Em750 Ex440/Em840 | | | | | | | | |
| | 4 | Ex310/Em330 Ex370/Em740 Ex500/Em790 Ex520/Em650 | | | | | | | | |
| | 5 | Ex300/Em430 Ex310/Em720 Ex340/Em760 Ex400/Em550 Ex470/Em510 | | | | | | | | |

To filter the results, all final pheromone trails of every model size (3 to 5), using *Ridge* and σ = 1, were normalized between 0 – 1 and summed. The hundred input variables with higher pheromone concentration were selected as the filtered inputs. In addition, the variables that participated in each global solution were added to this filter vector.

The AnTSbe was re-run, using the same general parameters as the first run, but only selecting variables within the filtered inputs. As for the model size, type, and base expansion, for this filtered run, we choose 3 – 6, *Ridge* regression and basis expansion 2, to evaluate if the expansion of the variables could improve the prediction metrics. Table 3.3 presents the metrics for this filtered run.

Table 3.3. Global solutions' metrics and selected fluorescence pairs for Diesel S100 - *filtered optimization* and PLS regression (7 LV) metrics for comparison.

| Diesel S100 Filtered *Ridge* and σ = 2 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Calibration | | | Validation | | | Test | | |
| | | $R^2$ | MAPE | RMSE | $R^2$ | MAPE | RMSE | $R^2$ | MAPE | RMSE |
| Model Size | 3 | 0.833 | 3.72 | 4.61 | 0.880 | 2.69 | 3.45 | 0.534 | 4.93 | 6.43 |
| | 4 | 0.923 | 2.63 | 3.12 | 0.824 | 3.75 | 4.17 | 0.648 | 4.68 | 5.59 |
| | 5 | 0.952 | 2.21 | 2.47 | 0.939 | 2.15 | 2.45 | 0.753 | 4.00 | 4.68 |
| | 6 | 0.953 | 2.13 | 2.43 | 0.947 | 1.98 | 2.29 | 0.75 | 4.20 | 4.72 |
| $PLS_7$ | | 0.990 | 0.83 | 1.12 | 0.332 | 6.30 | 8.13 | 0.39 | 6.51 | 7.38 |
| Selected Excitation/Emission Pairs | | | | | | | | | | |
| Model Size | 3 | Ex310/Em440 **Ex310/Em710** Ex440/Em840 | | | | | | | | |
| | 4 | **Ex270/Em470 Ex280/Em530 Ex310/Em710 Ex560/Em590** | | | | | | | | |
| | 5 | **Ex270/Em470 Ex280/Em530 Ex310/Em710 Ex500/Em790 Ex560/Em590** | | | | | | | | |
| | 6 | **Ex270/Em470 Ex280/Em530** Ex290/Em670 **Ex310/Em710 Ex500/Em790 Ex560/Em590** | | | | | | | | |

As can be seen, comparing the metrics in Table 3.2 and Table 3.3, the filtered run achieved better results than the previous optimizations. There was also a consensus throughout the internal runs of the algorithm about the selected pairs in each model size. Common pairs between different model sizes are bolded in Table 3.3. There was no significant improvement between selecting 5 or 6 input variables, so, by the principle of parsimony, there was no need to venture further into even bigger models.

As a direct comparison, linear models were built with classical partial least squares (PLS) regression using the same calibration, validation, and test subsets. Table 3.3 also presents the metrics for the PLS regression model using 7 latent variables (the number was selected by evaluating the minimum RMSE of prediction ranging from rank 1 to 20). The performance of the AnTSbe model with 5 inputs was significantly better than the PLS regression model for the validation and test subsets, requiring only a fraction of the information (PLS uses the whole spectra).

Figure 3.4 presents the measured *vs.* predicted outputs for the *Ridge, LassoLars,* and Filtered *Ridge* global solutions, with model size 5.

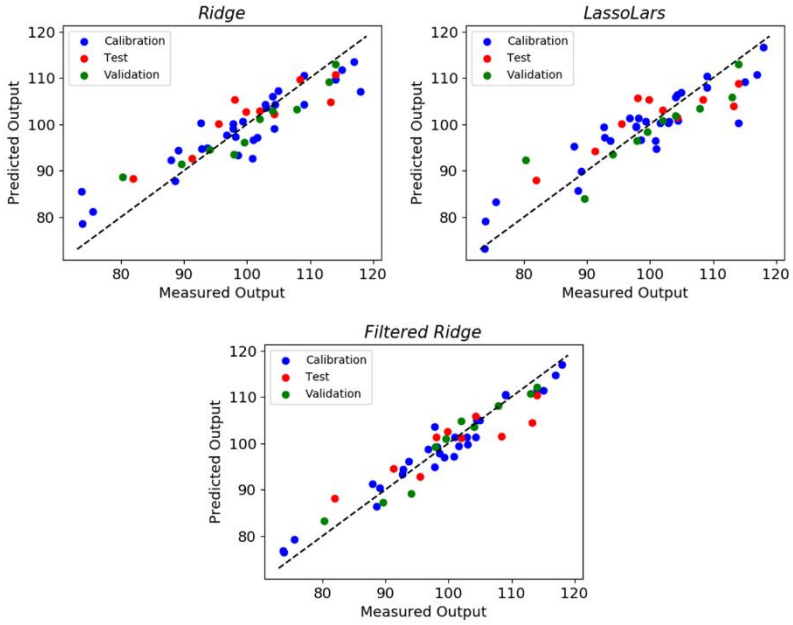



Figure 3.4. Diesel S100 Measured *vs* Predicted outputs for the *Ridge, LassoLars,* and Filtered *Ridge* global solutions, with model size 5.

For a more visual comprehension Figure 3.5 presents all the global solutions' (*Nw* = 5) selected fluorescence pairs plotted upon the average Diesel S100 EEM.

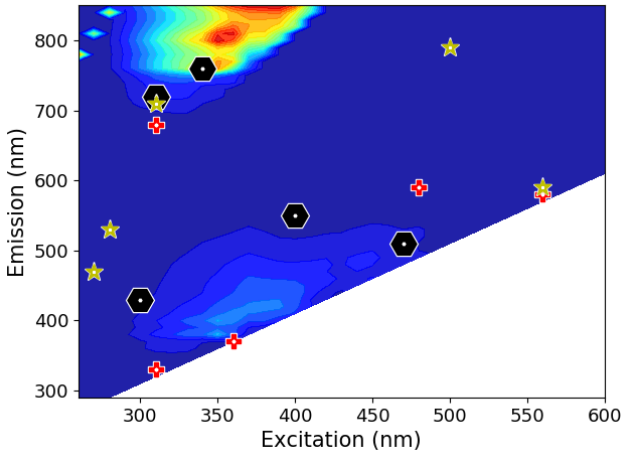



Figure 3.5. Global solutions' (*Nw* = 5) selected fluorescence pairs for Diesel S100 (Black hexagon – *LassoLars*; Red cross – *Ridge*; Yellow star – Filtered *Ridge*).

### *3.4.2 Diesel S10*

As both the first and the filtered optimization of the Diesel S10 models selected a combination of the same excitation/emission pairs, and the filtered run used the same optimization parameters, we will focus the results directly in the filtered global solutions.

Table 3.4 presents the global solutions' metrics and selected fluorescence pairs for the Diesel S10 filtered optimization.

Evaluating the global solutions, the predictive model using only 4 out of the original 1467 fluorescence pairs (less than 0.3%) is able to correctly quantify total sulfur in diesel S10 samples with considerable low errors. In addition, the validation and test metrics are similar to the calibration metrics, signalizing that the model could deal with unseen data

satisfactorily. There was no significant improvement between models with 4 and 5 input variables.

Table 3.4. Global solutions' metrics and selected fluorescence pairs for Diesel S10 - *filtered optimization*. PLS regression (4 LV) metrics for comparison.

| Diesel S10 Filtered *Ridge* and σ = 1 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Calibration | | | Validation | | | Test | | |
| | | $R^2$ | MAPE | RMSE | $R^2$ | MAPE | RMSE | $R^2$ | MAPE | RMSE |
| Model Size | 2 | 0.889 | 1.86 | 0.134 | 0.813 | 1.49 | 0.086 | 0.784 | 0.71 | 0.046 |
| | 3 | 0.952 | 1.25 | 0.088 | 0.968 | 0.62 | 0.036 | 0.750 | 0.72 | 0.050 |
| | 4 | 0.980 | 0.83 | 0.057 | 0.949 | 0.73 | 0.045 | 0.978 | 0.24 | 0.015 |
| | 5 | 0.986 | 0.71 | 0.048 | 0.994 | 0.26 | 0.016 | 0.991 | 0.15 | 0.009 |
| $PLS_4$ | | 0.997 | 0.12 | 0.008 | -8.58 | 9.22 | 0.619 | -5.12 | 4.19 | 0.247 |
| Selected Excitation/Emission Pairs | | | | | | | | | | |
| Model Size | 2 | Ex310/Em640 **Ex400/Em760** | | | | | | | | |
| | 3 | **Ex260/Em790 Ex340/Em810** Ex390/Em570 | | | | | | | | |
| | 4 | **Ex340/Em810** Ex380/Em450 Ex520/Em650 Ex600/Em740 | | | | | | | | |
| | 5 | **Ex260/Em790** Ex290/Em450 **Ex340/Em810 Ex400/Em760** Ex410/Em660 | | | | | | | | |

As for Diesel S100, PLS regression models were also built using Diesel S10 whole spectral data and the same calibration, validation, and test subsets, and the model metrics can be seen in Table 3.4. Comparing the AnTSbe and the PLS models, we can see that the later had metrics one order of magnitude higher than the former. Although non-intuitive, $R^2$ (as defined in Equation 5) can assume values between $-\infty$ and 1 (O. KVÅLSETH, 2012). This can happen when the fitted model has predictions that are worse than a horizontal line equal to the mean value of the output in that subset. Generally, $R^2$ is nonnegative for any linear regression with intercept, and that will always occur in the calibration subset. When using the fitted model to predict the validation and test subset, if the model is overfitted or unable to deal with this unseen data, its predictions can be actually worse than a constant (equal to the mean value of the output). When $R^2$ is negative, it indicates a complete lack of fit. In the Diesel S10 case, there is a small amount of validation/test samples, and their total sulfur concentration has a very small amplitude within the group. This way, the predictive model must have very small prediction errors to be better than the group means. The AnTSbe models achieved the necessary accuracy, but not the PLS model.

Figure 3.6 presents the measured *vs.* predicted outputs for the Filtered *Ridge* global solution, with model size 4.

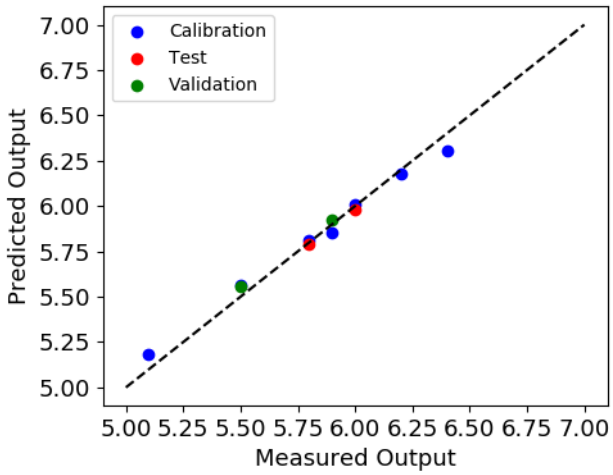



Figure 3.6. Diesel S10 Measured *vs.* Predicted outputs for the Filtered *Ridge* global solution, with model size 4.

Finally, Figure 3.7 shows the global solutions' (*Nw* = 4) selected fluorescence pairs plotted upon the average Diesel S10 EEM.

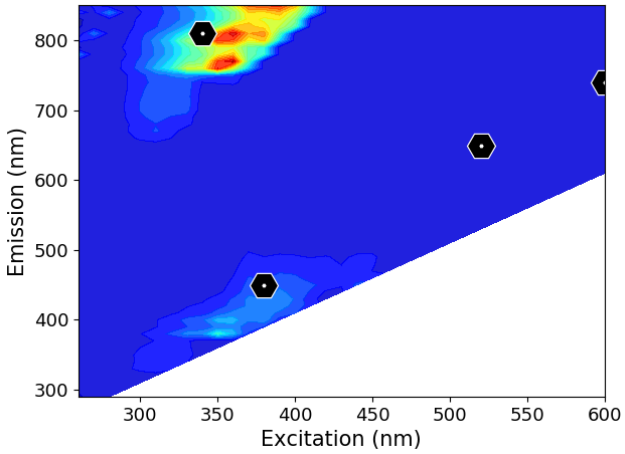



Figure 3.7. Global solution' (*Nw* = 4) selected fluorescence pairs for Diesel S10.

Analyzing Figure 3.5 and Figure 3.7, we can see that both local models selected fluorescence pairs in regions that are far from the peak of fluorescence intensity. This can be explained by remembering that the black-box models fitted here use empirical data without any need for phenomenological significance. Even though those pairs appear to be in a region of noise and no fluorescence, after normalization, they follow linear trends that can be correlated to sulfur in the samples. The algorithm treats all pairs equally, no matter their relative intensity, and seeks the model with the smallest errors, as can be seen by the metrics presented. Pairs, especially the ones between Ex 260 to 400 and Em 300 to 500, could be correlated to the fluorescence of benzothiophenes and dibenzothiophenes observed in other works (AARON et al., 2002; BREE; ZWARICH, 1971; HOU et al., 2019; NAYAK; AGARWAL; PERIASAMY, 2010). However, those same works state that the differences in solvents and radicals attached to the compounds can shift the fluorescence peak to other regions. This way, it is hard to directly link fluorescence pairs to sulfur-containing molecules based only on the fitted models. For future works, if the chemical meaning of the fluorescence pairs is relevant, it is possible to spike diesel fuel samples with known diesel sulfur-containing compounds and evaluate models based on this controlled

changes. Also, the valid region to select features of the spectra can be trimmed to known areas with high fluorescence intensity.

### *3.4.3 Tabu Memory Activations*

To study the impact of the tabu memory list, we follow how many times, on average, a forbidden combination was chosen by the ants throughout all performed optimizations. As expected, some parameters had a direct correlation with tabu memory activations. First, the number and source of inputs matter: the higher the number of inputs, the rarer will be the situation of the stochastic selection of the same elements by ants. The source is also relevant. In cases like the one presented (using EEM fluorescence), many neighbor inputs carry similar information. In cases where inputs are more linear-independent, the activations would be more frequent. Other very influential parameter is the model size. Smaller models have more chances to activate de tabu memory, and that is corroborated both by the number of possible combination of inputs by permutation (that considerably increases with model size) as by the way the pheromone vector is constructed: the pheromone distribution is uneven in the top performers and more equilibrated in the other inputs. This can be seen in Figure 3.8, which illustrates the individually normalized final pheromone trail of optimizations with diesel S100 and model sizes 3, 4, and 5.

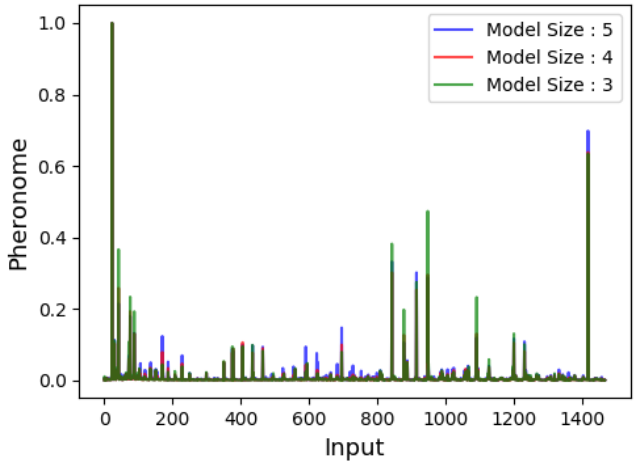



Figure 3.8. Individually normalized final pheromone trail of optimizations with Diesel S100 and model sizes 3, 4, and 5.

After the two first inputs with higher pheromone concentration (inputs 23 and 1416), there is a close competition for the subsequent top performers. This balance in pheromone distribution reflects in the input selection by the ants, meaning that bigger models have smaller chances of selecting forbidden combinations.

For the number of activations of the Tabu memory, in the first optimizations (1467 inputs), model sizes 3, 4, and 5 had 500, 100, and 10 activations, respectively, for diesel S10. Diesel S100 had in general, only a fifth of that – 100, 20, ~1. This difference is due to the higher number of fluorescent-sulfured components present in diesel S100, making the information contained in the spectra more disperse. In the filtered runs (100 inputs), both diesel S10 and S100 had, on average, 6.000, 2.000, and 200 activations for model sizes 3, 4, and 5. Considering that each run fits 30.000 models (200 ants and 150 iterations), 6.000

activations are equivalent to repeating 20% of all fitted models. For the reasons discussed above, in this case study, the tabu memory had no influence in model sizes 5 or bigger.

### *3.4.4 Contrast with Previous Works*

Our research group also studied the application of fluorescence spectroscopy as a tool to predict sulfur content in diesel fuel in a previous work (RANZAN et al., 2015). In it, the chosen optimization strategy applied for the selection of fluorescence pairs and the fitting of linear models was called Pure Spectra Chemometric Modelling (PSCM): an ACO based algorithm developed for the selection of spectral elements to predict state variables. The AnTSbe can be seen as an evolution of the PSCM, but for general use. The major differences between algorithms are the intended contributions of this work: the introduction of regularized linear models, hybridization with Tabu Search, the adjustable mechanisms of pheromone manipulation, the use of polynomial input expansion, and filtration runs. Both works used the same diesel S10 and S100 datasets, being the only difference that the PSCM diesel S100 group had one more sample, containing 138 ppm of total sulfur. In addition, the splitting of the data was not the same: The PSCM study applied systematic sampling to split the datasets into only training and test subsets, in a 2:1 proportion. Disregarding these small differences, the results of both works can be reasonably compared. Table 3.5 presents a compilation of the Global Solutions of both methodologies for diesel S100 and S10, with model sizes 5 and 4, respectively. As can be seen in Table 3.5, the proposed modifications made a positive impact in the general quality of the final predictive models.

Table 3.5. Comparison between Global Solutions of AnTSbe and PSCM (RANZAN et al., 2015) methodologies for diesel S100 and S10, with model sizes 5 and 4, respectively.

| **S100** | Calibration | | Validation | | Test | |
|---|---|---|---|---|---|---|
| | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE |
| $AnTSbe_5$[a] | 0.952 | 2.47 | 0.934 | 2.45 | 0.753 | 4.68 |
| $PSCM_5$ | 0.660 | 7.41 | - - - | - - - | 0.410 | 8.94 |
| **S10** | Calibration | | Validation | | Test | |
| | $R^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE |
| $AnTSbe_4$[b] | 0.980 | 0.057 | 0.949 | 0.045 | 0.978 | 0.015 |
| $PSCM_4$ | 0.940 | 0.09 | - - - | - - - | 0.690 | 0.16 |

[a] Filtered *Ridge* $\sigma = 2$ and [b] Filtered *Ridge* $\sigma = 1$

In both studies, the same Diesel S10 dataset was used as a case to evaluate and present the evolution in the methodologies. However, the small number of samples must be taken into account when applying the predictive models to other researches or industrial applications. The authors strongly recommend that future works that intend to predict sulfur concentration in ultra-low-sulfur diesel increase the size of the dataset to reparametrize the presented models or even rerun the algorithm in search of more suitable case-specific models.

## 3.5 Conclusions

Fluorescence EEM is a fast and viable source of information that can be correlated to many properties in diesel fuel. Applying the AnTSbe methodology, we were able to optimize the selection of input variables and fit predictive models that use a small amount of fluorescence data to estimate total sulfur concentration in diesel samples.

For the Diesel S100 local models, using only 5 out of the original 1467 fluorescence pairs, combined with basis expansion, we were able to satisfactorily predict total sulfur content in samples with mean absolute prediction errors of less than 4% and root mean squared errors of 4.68 ppm, for the test subset. For the Diesel S10 local models, the use of 4 Ex/Em pairs was sufficient to predict sulfur content with MAPE 0.24%, and RMSE of 0.015 ppm, for the test subset.

Comparing the AnTSbe global solutions with classical PLS regression models and previous optimization studies (PSCM), the proposed methodology proved superior in the task of predicting sulfur content in real diesel fuel samples without the need for any physical or chemical pretreatment. Furthermore, the proposed AnTSbe methodology can deal not only with fluorescence data, but could be used for optimizations with any source of information, as infrared/Raman spectrum, industrial measurements, physicochemical properties, or even a combination of those. The introduction of the tabu memory list was useful to avoid early stagnation and redundant calculations, especially in models with smaller sizes. The technique could be even more fruitful in studies where the input variables are less correlated. The use of basis expansion also proved beneficial for the predictive models.

Using only a small selected fraction of the original spectra, we reduce the time required to acquire data, the influence of noisy inputs that do not carry valid information, and simplify the equipment specification. The proposed methodology is a further step for the construction of custom sensors that could be coupled directly into refinery streams to predict sulfur or even other compounds that present natural fluorescence. This predictive information can be used by operators to take effective controlling actions in hydrodesulphurization processes.

## 3.6 Acknowledgment

The authors are grateful for the scholarship provided by CAPES and would like to thank Petrobras for financial and technical support.

## 3.7 Abbreviations

ACO, Ant Colony Optimization; Ex/Em, excitation/emission; EEM, excitation-emission matrix; HDS, hydrodesulfurization; LV, latent variable; MAPE, mean absolute prediction error; PLS, Partial Least Squares; PCA, Principal Component Analysis; PSCM, Pure Spectra Chemometric Modelling; RMSE, root mean squared error; S10, sample group with less than 10 ppm of sulfur; S100, sample group with average 100 ppm of sulfur; TS, Tabu Search.

## 3.8 References

AARON, J. J. et al. **Luminescence Properties of New Fused Benzothiophene Derivatives and Their Conductive Oligomers Structural and Solvent Effects**. Journal of Fluorescence. **Anais**...Springer, jun. 2002Disponível em: <https://link.springer.com/article/10.1023/A:1016869002735>. Acesso em: 28 jul. 2020

ABDULKADER, M. M. S.; GAJPAL, Y.; ELMEKKAWY, T. Y. Hybridized ant colony algorithm for the Multi Compartment Vehicle Routing Problem. **Applied Soft Computing**, v. 37, p. 196–203, 2015.

ABURTO, P. et al. Quantitative analysis of sulfur in diesel by enzymatic oxidation, steady-state fluorescence, and linear regression analysis. **Energy and Fuels**, v. 28, n. 1, p. 403–408, 2014.

ARITO, F.; LEGUIZAMÓN, G. **Incorporating Tabu Search Principles into ACO Algorithms**. [s.l: s.n.]. v. 5818

ASTM. **D4294-10 Standard Test Method for Sulfur in Petroleum and Petroleum Products by Energy Dispersive X-ray Fluorescence Spectrometry**West Conshohocken, PA, 2010.

ASTM. **D7039-15a Standard Test Method for Sulfur in Gasoline, Diesel Fuel, Jet Fuel, Kerosine, Biodiesel, Biodiesel Blends, and Gasoline-Ethanol Blends by Monochromatic Wavelength Dispersive X-ray Fluorescence Spectrometry**West Conshohocken, PA, 2015.

BALSEIRO, S.; LOISEAU, I.; RAMONET, J. An Ant Colony algorithm hybridized with insertion heuristics for the Time Dependent Vehicle Routing Problem with Time Windows. **Computers & OR**, v. 38, p. 954–966, 2011.

BELTRAMO, T.; KLOCKE, M.; HITZMANN, B. Prediction of the biogas production using GA and ACO input features selection method for ANN model. **Information Processing in Agriculture**, v. 6, n. 3, p. 349–356, 2019.

BLUM, C. et al. **Hybrid Metaheuristics: An Emerging Approach to Optimization**. [s.l: s.n.].

BLUM, C. et al. Hybrid metaheuristics in combinatorial optimization: A survey. **Applied Soft Computing**, v. 11, n. 6, p. 4135–4151, 2011.

BREE, A.; ZWARICH, R. Electronic spectra of dibenzothiophene. **Spectrochimica Acta Part A: Molecular Spectroscopy**, v. 27, n. 4, p. 621–630, 1 abr. 1971.

BRUNET, S. et al. On the hydrodesulfurization of FCC gasoline: a review. **Applied Catalysis A: General**, v. 278, n. 2, p. 143–172, 2005.

BULLNHEIMER, B.; HARTL, R.; STRAUSS, C. A New Rank Based Version of the Ant System - A Computational Study. **Central European Journal of Operations Research**, v. 7, p. 25–38, 1999.

CAMACHO, J.; PICÓ, J.; FERRER, A. Data understanding with PCA: Structural and Variance Information plots. **Chemometrics and Intelligent Laboratory Systems**, v. 100, n. 1, p. 48–56, 2010.

CAMPOS, A. T. et al. Prediction of Sulfur Content in Diesel/Biodiesel Blends Using LED-Induced Fluorescence Associated with Multivariate Calibration. **Journal of the Brazilian Chemical Society**, v. 29, p. 1367–1372, 2018.

CHANDRA MOHAN, B.; BASKARAN, R. A survey: Ant Colony Optimization based recent research and implementation on several engineering domain. **Expert Systems with Applications**, v. 39, n. 4, p. 4618–4627, 2012.

CHANDRA SRIVASTAVA, V. An evaluation of desulfurization technologies for sulfur removal from liquid fuels. **RSC Advances**, v. 2, n. 3, p. 759–783, 2012.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers and Electrical Engineering**, v. 40, n. 1, p. 16–28, 1 jan. 2014.

DORIGO, M.; BLUM, C. Ant colony optimization theory: A survey. **Theoretical Computer Science**, v. 344, n. 2–3, p. 243–278, 2005.

DORIGO, M.; GAMBARDELLA, L. M. Ant colonies for the travelling salesman problem. **Biosystems**, v. 43, n. 2, p. 73–81, 1997.

EFRON, B. et al. Least angle regression. **The Annals of Statistics**, v. 32, n. 2, p. 407–499, 2004.

EINIPOUR, A.; CORRESPONDING. A fuzzy-ACO method for detect breast cancer. **Global Journal of Health Science**, v. 3, 2011.

GAMBARDELLA, L.; DORIGO, M. Ant-Q: A Reinforcement Learning approach to the traveling salesman problem. **Proceedings of ML-95, the 12th International Conference on Machine Learning**, v. 170, 2000.

GODOY, J. L.; VEGA, J. R.; MARCHETTI, J. L. Relationships between PCA and PLS-regression. **Chemometrics and Intelligent Laboratory Systems**, v. 130, p. 182–191, 15 jan. 2014.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization and Machine Learning**. Boston: Kluwer Academic Publishers, 1989.

GUIDOTTI, R. et al. A Survey of Methods for Explaining Black Box Models. **ACM Computing Surveys**, v. 51, 2018.

HOU, Y. et al. Effects of polycyclic aromatic hydrocarbons on the UV-induced fluorescence spectra of crude oil films on the sea surface. **Marine Pollution Bulletin**, v. 146, p. 977–984, 1 set. 2019.

HU, L. et al. Vis-NIR spectroscopy Combined with Wavelengths Selection by PSO Optimization Algorithm for Simultaneous Determination of Four Quality Parameters and Classification of Soy Sauce. **Food Analytical Methods**, v. 12, n. 3, p. 633–643, 2019.

HUA, R. et al. Determination of sulfur-containing compounds in diesel oils by comprehensive two-dimensional gas chromatography with a sulfur chemiluminescence detector. **Journal of Chromatography A**, v. 1019, n. 1–2, p. 101–109, 2003.

JOHNSON, B.; XIE, Z. Classifying a high resolution image of an urban area using super-object information. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 83, p. 40–49, 2013.

KAVZOGLU, T. Chapter 33 - Object-Oriented Random Forest for High Resolution Land Cover Mapping Using Quickbird-2 Imagery. In: SAMUI, P.; SEKHAR, S.; BALAS, V. E. (Eds.). . **Handbook of Neural Computation**. [s.l.] Academic Press, 2017. p. 607–619.

KLUABWANG, J.; PUANGDOWNREONG, D.; SUJITJORN, S. Multipath Adaptive Tabu Search for a Vehicle Control Problem. **Journal of Applied Mathematics**, v. 2012, 2012.

KU-MAHAMUD, K.; ALOBAEDY, M. **New Heuristic Function in Ant Colony System Algorithm for Optimization**. [s.l: s.n.].

LI, X. Y.; TIAN, P.; LEUNG, S. C. H. An ant colony optimization metaheuristic hybridized with tabu search for open vehicle routing problems. **Journal of the Operational Research Society**, v. 60, n. 7, p. 1012–1025, 2009.

LIMA, F. et al. Towards a sulfur clean fuel: Deep extraction of thiophene and dibenzothiophene using polyethylene glycol-based deep eutectic solvents. **Fuel**, v. 234, p. 414–421, 2018.

MANSOURI, M. et al. **Statistical Fault Detection of Chemical Process - Comparative Studies**. [s.l: s.n.]. v. 07

MASSARON, L. AUTHOR; BOSCHETTI, A. AUTHOR C. N.-B. L. H. D. R. T. E. D. 8770. **Regression Analysis with Python**. Birmingham B3 2PB, UK: Packt Publishing Ltd., 2016.

MOHAPATRA, S.; PATRA, D.; SATPATHY, S. An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. **Neural Computing and Applications**, v. 24, n. 7, p. 1887–1904, 2014.

MORO, S.; RITA, P.; VALA, B. Predicting social media performance metrics and evaluation

of the impact on brand building: A data mining approach. **Journal of Business Research**, v. 69, n. 9, p. 3341–3351, 2016.

MURAKAMI, Y. Analysis of corrosive wear of diesel engines: relationship to sulfate ion concentrations in blowby and crankcase oil. **JSAE Review**, v. 16, n. 1, p. 43–48, 1995.

MYSZKOWSKI, P. B. et al. Hybrid ant colony optimization in solving multi-skill resource-constrained project scheduling problem. **Soft Computing**, v. 19, n. 12, p. 3599–3619, 2015.

NAYAK, P. K.; AGARWAL, N.; PERIASAMY, N. **Synthesis, photophysical and electrochemical properties of 2,8-diaryl-dibenzothiophene derivatives for organic electronicsJ. Chem. Sci**. [s.l: s.n.].

NIU, Y. et al. A Hybrid Tabu Search Algorithm for a Real-World Open Vehicle Routing Problem Involving Fuel Consumption Constraints. **Complexity**, v. 2018, p. 1–12, 2018.

O. KVÅLSETH, T. Cautionary Note About R2. **The American Statistician**, v. 39, p. 279–285, 2012.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, n. Oct, p. 2825–2830, 2011.

PESSOA, C. M. et al. Development of Ant Colony Optimization (ACO) Algorithms Based on Statistical Analysis and Hypothesis Testing for Variable Selection. **IFAC-PapersOnLine**, v. 48, n. 8, p. 900–905, 2015.

PIRIM, H.; BAYRAKTAR, E.; EKSIOGLU, B. Tabu Search: A Comparative Study. In: **Tabu Search**. [s.l.] I-Tech Education and Publishing, 2008.

RANZAN, C. et al. Wheat flour characterization using NIR and spectral filter based on Ant Colony Optimization. **Chemometrics and Intelligent Laboratory Systems**, v. 132, n. 0, p. 133–140, 2014.

RANZAN, C. et al. Sulfur Determination in Diesel using 2D Fluorescence Spectroscopy and Linear Models. **IFAC-PapersOnLine**, v. 48, n. 8, p. 415–420, 2015.

RANZAN, L. et al. Classification of Diesel Fuel Using Two-Dimensional Fluorescence Spectroscopy. **Energy & Fuels**, v. 31, n. 9, p. 8942–8950, 2017.

RASCHKA, S. AUTHOR C. N.-006. 3. 23 B. L. H. D. R. T. E. D. 8778. **Python machine learning**. Birmingham B3 2PB, UK.: Packt Publishing Ltd., 2015.

RUSSELL, E. L.; CHIANG, L. H.; BRAATZ, R. D. Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 51, n. 1, p. 81–93, 2000.

SAGBAN, R.; KU-MAHAMUD, K.; ABU BAKAR, M. **Reactive memory model for ant colony optimization and its application to TSP**. [s.l: s.n.].

SANTOS, P. V. J. L. et al. K-RANK: AN EVOLUTION OF Y-RANK FOR MULTIPLE SOLUTIONS PROBLEM. **Brazilian Journal of Chemical Engineering**, v. 36, p. 409–419, 2019.

SEBBEN, J. A. et al. Development of a quantitative approach using Raman spectroscopy for carotenoids determination in processed sweet potato. **Food Chemistry**, v. 245, p. 1224–1231, 2018.

SERBENCU, A.; MINZU, V. **Hybridized Ant Colony System for Tasks to Workstations Assignment**. 2016 IEEE Symposium Series on Computational Intelligence (SSCI). **Anais**...2016

SILVA, A. C. DA et al. Two-dimensional linear discriminant analysis for classification of three-way chemical data. **Analytica Chimica Acta**, v. 938, p. 53–62, 2016.

SILVA, J. I. S. DA; SECCHI, A. R. AN APPROACH TO OPTIMIZE COSTS DURING ULTRA-LOW HYDRODESULFURIZATION OF A BLEND CONSISTING OF DIFFERENT OIL STREAMS. **Brazilian Journal of Chemical Engineering**, v. 35, p. 1293–1304, 2018.

STÜTZLE, T.; LÓPEZ-IBÁÑEZ, M.; DORIGO, M. A Concise Overview of Applications of Ant Colony Optimization. In: [s.l: s.n.].

SUN, L. et al. A Hybrid Gene Selection Method Based on ReliefF and Ant Colony Optimization Algorithm for Tumor Classification. **Scientific Reports**, v. 9, n. 1, 2019.

TANG, J.; ALELYANI, S.; LIU, H. Feature selection for classification: A review. In: **Data Classification: Algorithms and Applications**. [s.l: s.n.]. p. 37–64.

TIKHONOV, A. N.; ARSENIN, V. I. A. **Solutions of ill-posed problems**. [s.l.] Winston, 1977.

TIPPING, M. E.; BISHOP, C. M. Probabilistic Principal Component Analysis. **Journal of the Royal Statistical Society. Series B (Statistical Methodology)**, v. 61, n. 3, p. 611–622, 1999.

TSUTSUI, S.; FUJIMOTO, N. **ACO with Tabu Search on a GPU for Solving QAPs using Move-Cost Adjusted Thread Assignment**. [s.l: s.n.].

WANG, F. C. Y. et al. Speciation of Sulfur-Containing Compounds in Diesel by Comprehensive Two-Dimensional Gas Chromatography. **Journal of Chromatographic Science**, v. 41, n. 10, p. 519–523, 2003.

WOLD, S.; SJÖSTRÖM, M.; ERIKSSON, L. PLS-regression: a basic tool of chemometrics. **Chemometrics and Intelligent Laboratory Systems**, v. 58, n. 2, p. 109–130, 2001.

ZHANG, P. (ED.). Front Matter. In: **Advanced Industrial Control Technology**. Oxford: William Andrew Publishing, 2010. p. iii.

ZHANG, X.; TANG, L. **A New Hybrid Ant Colony Optimization Algorithm for the Traveling Salesman Problem**. (D.-S. Huang et al., Eds.)Advanced Intelligent Computing Theories and Applications. With Aspects of Artificial Intelligence. **Anais**...Berlin, Heidelberg: Springer Berlin Heidelberg, 2008

ZHENG, C. et al. Feature selection for high-dimensional integrated data. In: **Proceedings of the 2012 SIAM International Conference on Data Mining**. Proceedings. [s.l.] Society for Industrial and Applied Mathematics, 2012. p. 1141–1150.

# Capítulo 4 – **Prediction of Total Phenolic Content in Wood-Aged Cachaças using a Hybrid Ant Colony – Tabu Search algorithm and Fluorescence Spectroscopy with a Reference Spectral Pair**

**Abstract**: Fluorescence spectroscopy is a non-invasive and non-destructive technique widely applied in identifying and quantifying natural fluorophores. As an auxiliary measure, there is a need to use chemometric tools to extract useful information from the spectra. In this study, we propose an adaptation to a hybrid metaheuristic optimization algorithm, especially design for fluorescence data, introducing the concept of a Delta Pair (ΔP), which will act as a reference in the spectra. The selection of this fluorescence pair is made by adding an external feature-selecting layer based on Ant Colony Optimization to the routine. To evaluate the proposed methodology, we will predict total phenolic concentration applying the adapted algorithm to fluorescence data of three *cachaças* (sugarcane spirit) aged in Amburana wood barrels in our laboratory and commercially aged *cachaças*. Following the fluorescence profile of the aging cachaças, it is possible to observe a displacement of the fluorescence peak and suppress fluorescence bands as the phenolic content in the spirit increases. In the global model using all available samples (MD – models using $CA_1$/$CA_2$/$CA_3$ data) and polynomial combination of inputs, the use of the ΔP reduced the test RMSE and MAPE from 74.60 mg EAG $L^{-1}$ and 39.7 % (original) to 43.9 mg EAG $L^{-1}$ and 18.9 % (adapted). The same trend was seen when fitting local models (AD – models using $CA_2$ and $CA_3$ data). The use of the ΔP reduced test MAPE from 36.2% to 13%. The results shown corroborate with the proposed methodology and support that the Delta Pair adaptation was necessary to improve the generalization capabilities of the models.

## 4.1 Introduction

Excitation-Emission-Matrix (EEM) fluorescence spectroscopy is a powerful technique for the simultaneous identification and quantification of many fluorophores (FAASSEN; HITZMANN, 2015). Being fast, non-destructive, and non-invasive, it is especially praised for bioprocess applications, where contamination can be disastrous (BECKER et al., 2007). As an auxiliary measurement, there is a need to use chemometric tools to extract useful information from the spectra (KUMAR et al., 2014). Commonly applied methodologies can involve classical Principal Component Analysis – PCA and Partial Least Squares – PLS (KETTANEH; BERGLUND; WOLD, 2005), PARAFAC/Tucker (SILVA et al., 2019), and Convolutional Neural Networks (ITAKURA et al., 2018), but all these methods require the use of the whole fluorescence spectra. To simplify data acquisition and discard unnecessary information, researchers also apply feature selection to EEM fluorescence spectra to optimize selecting only the Excitation/Emission pairs that aggregate valid information to solve their specific problem (ASSAWAJARUWAN; REINALTER; HITZMANN, 2017; ÖDMAN et al., 2009). Ranzan *et al.* (2020) proposed a hybrid metaheuristic ant colony – tabu search algorithm with polynomial combinational of inputs to optimize linear models (called AnTSbe) that could be applied to fluorescence data. The problem with higher-order variables is that they potentialize the interference of contaminants and have smaller generalization capabilities. In this contribution, we propose an adaptation of the AnTSbe algorithm, especially design for fluorescence data, introducing the concept of a Delta Pair, which will serve as a reference in the spectra. The Delta Pair will be chosen by adding another feature selection layer to the AnTSbe, with its independent pheromone trail. At the end of the optimization routine, the favored Delta Pair will be the one that, with its fluorescence intensity decreased from all other Ex/Em pairs in the spectra, resulted in the best fitted predictive model. To evaluate the proposed methodology, we will predict total phenolic concentration applying the adapted algorithm to EEM fluorescence data of three *cachaças* aged in wood barrels in our laboratory and commercial, aged cachaças a continuation of the studies initiated by Carvalho *et al.* (2020). Brazil is the world's biggest cachaça producer, with an installed production capacity of 1.2 billion liters per year (IBRAC, 2019). Although essential to incorporate value to the final product, aging processes still rely on specialized tasting personal and static, pre-defined aging periods (MOSEDALE; PUECH, 1998). As many of the sensorial changes in the aged spirit can be related to phenolic compounds, the rapid and non-invasive characterization of the aging profile and the prediction of these compounds can be valid tools for the modernization of the industry.

## 4.2 Material and Methods

### *4.2.1 Barrels and Cachaça for Aging*

For this study, three cachaça aging processes were performed in our laboratory and will be referred to as CA1, CA2, and CA3. Each aging had particular characteristics, but common to all, the barrels were placed in a dark cabinet, protected from vibrations and sudden temperature changes. During the aging process, samples were regularly collected from the barrels and stored in amber bottles, protected from light, and conditioned under freezing conditions (-18°C). The EEM fluorescence spectra of each sample were collected,

and the concentration of total phenolic compounds was quantified by the adapted Folin-Ciocalteau method (described below).

**CA1:** Cachaçaria Weber Haus (Ivoti/RS) provided the premium cachaça used in this aging. It had 48% of ethanol (v/v), higher than the usual 38 - 40% commercial concentration, and was acquired directly in the distillery. The cachaça was aged in a virgin, lightly toasted amburana (native Brazilian tree) barrel of 20 liters for a total of 463 days. The study of this first aging can be seen in Carvalho *et al.* [10] and was one of the inspirations of this work. The aged cachaça was bottled, and the barrel was gently washed with non-aged cachaça to prepare for CA2 process.

**CA2:** In this aging, commercial cachaça was used. The chosen brand was 'Pirassununga 51', a famous Brazilian cachaça, with 39% ethanol (v/v) concentration. The cachaça was aged in the same amburana barrel from CA1 for a total of 205 days.

**CA3:** The cachaça used was also commercial Pirassununga 51 (from the same manufacturing lot as CA2). The barrel used was a new 20-liter virgin amburana barrel. It was produced by a different cooperage than the CA1/CA2 barrel. It was coated internally with paraffin (a standard procedure adopted by some cooperages to avoid leakage in low-end products). Before filling the barrel with cachaça, it was repeatedly washed with boiling water to remove the paraffin. This washing routine also affected the internal roasted layer of the wood, removing a fair amount of color and debris, which are beneficial for the aging process. The cachaça was aged for 205 days.

Along with our laboratory-aged cachaça, eight commercially aged cachaças were also evaluated. Three were aged in amburana for one year – Amb01, Amb02, and Amb03; one aged in balsam for one year – Balm01; one aged five years in oak and then one year in balsam – Mix01; one aged for two years in a blend of seven kinds of wood (French oak, American oak, balsam, amburana, garapa (*Apulela leiocarpa*), cabreúva (*Myrocarpus frondosus*) and Brazilian sassafras) – Blend01; one aged in oak – Oak01 and one aged in cabreúva – Cab01, the last two brands without information of aging period.

### *4.2.2 Quantification of Total Phenolic Concentration*

The quantification of total phenolic compounds was performed by the adapted spectrophotometric method of Folin-Ciocalteau (SINGLETON; ROSSI; JR, 1985) described in Carvalho *et al.* (2020). The samples were diluted with ultrapure water (Milli-Q system), mixed with Folin-Ciocalteau 2N and $Na_2CO_3$ 7% (w/v), and rested for 1 hour at 25°C, protected from light; the absorbance was measured at 765 nm in UV-VIS Spectrophotometer (1600A, Pro-Análise). The quantification of phenolic compounds was performed using a standard curve of gallic acid. The results were expressed as milligrams of gallic acid equivalent per liter of a sample (mg GAE $L^{-1}$). All analyses were performed in triplicate.

### *4.2.3 EEM Fluorescence Spectroscopy*

The EEM fluorescence spectra were collected using a Fluoromax-4 spectrofluorometer (Horiba, Japan), equipped with a xenon lamp of 150 W. The measurements were made in the range of excitation wavelengths between 260 and 600 nm and emission wavelengths between 290 and 850 nm, with a resolution of 10 nm. The geometry of the measurements was 90°, and all measurements were made in triplicate. Each fluorescence spectrum

obtained with these arrangements was a 57x35 matrix containing the fluorescence intensity of 1995 excitation/emission pairs (Ex/Em). As no excitation can lead to emission with higher energy, pairs where Em < Ex were discarded. First and second-order Rayleigh scattering were removed (BAHRAM et al., 2006) once they were very prominent in samples with few aging days. The chosen approach was to replace the scattering areas with missing values, which will not be considered during variable selection. Each EEM (relative to each sample) was unfolded into a vector of size 1 × 1065, the row relating to the sample and the columns containing the fluorescence intensity of each valid Ex/Em pair, and vertically stacked.

### *4.2.4 Data Preprocessing*

The fluorescence EEM data preprocessing started with the splitting into calibration, validation, and test subsets. The applied methodology was *K-rank* (SANTOS et al., 2019), sorting the whole data in ascending phenolic concentration and adapting a pattern to select, in order, samples to their respective subsets. In this case, 60% to calibration, 20% to validation, and 20% to test. The *StandardScaler* (preprocessing function of the Python scikit-learn library (PEDREGOSA et al., 2011)) was used to scale the data to mean zero and unit variance (independently in each feature). The Scaler is fitted using the calibration subset and then applied to the validation and test subsets.

### *4.2.5 Chemometric analysis – Adapted AnTSbe algorithm*

AnTSbe is a compilation of tools to apply stochastic variable selection and linear model optimization to multidimensional data. The algorithm is based on Ant Colony Optimization, a stochastic optimization algorithm where multiple parallel processes (ants) take different 'routes' to minimize an objective function. The ants ascribe quality indicators, called pheromones, to the input variables. The pheromone of each input is incremented after each iteration based on how well a model using this variable could predict the desired output. At each new iteration, features are selected by the ants using the pheromone trail and a random trigger, being able to wander through any possible solution but with a higher chance to select inputs with a greater concentration of pheromones. The algorithm is hybridized with Tabu Search, a neighborhood search-based method, which uses a memory structure to avoid being trapped in local optima. The AnTSbe keeps a tabu memory list to avoid previously tested input combinations. The tabu memory is updated at every iteration, adding the newly tested and forgetting combinations from older iterations. The complete methodology can be seen in Ranzan *et al.* (2020).

The AnTSbe is divided into three phases. *Phase One* is the initialization stage, where the optimization parameters are chosen. The parameters are model size (number of variables in the linear model), type of model (*e.g.,* Ridge Regression and Lasso Regression), base expansion (polynomial combinations of features with degree equal or less than σ), optimization metric (MAPE or RMSE), number of runs, number of iterations, number of ants, tabu memory size (keep tabu combinations for how many iterations), initial pheromone value, pheromone gain, and pheromone evaporation rate. *Phase Two* is the core of the optimization, where models are fitted, evaluated, and compared. In *Phase Three*, the best-fitted models are presented, with their corresponding variables and all

comparative metrics. A schematic representation of the algorithm can be seen in Figure 4.1 (A).

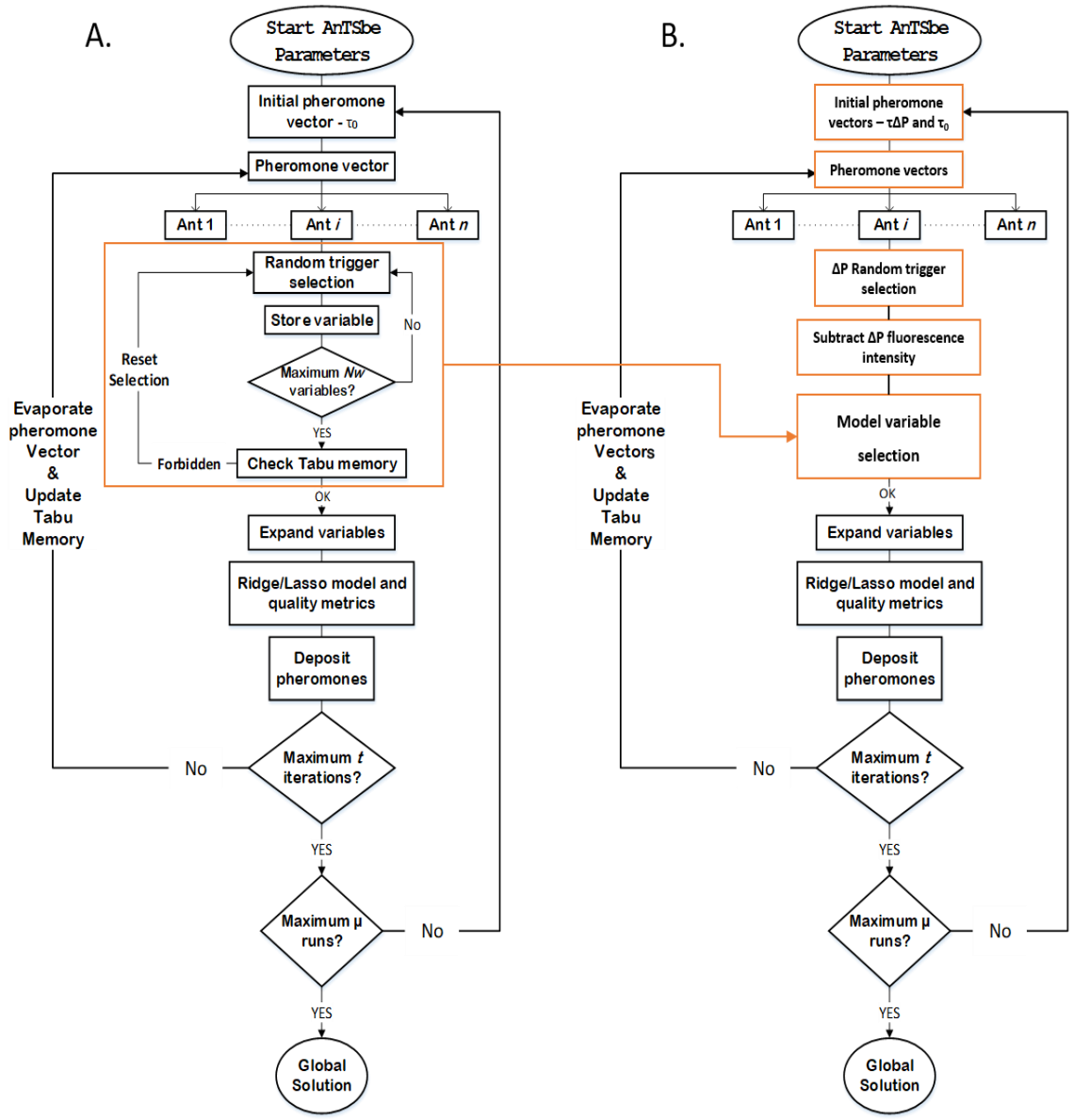



Figure 4.1. Schematic representation of the original AnTSbe algorithm (A) and the proposed adapted AnTSbe (B).

The first proposal of this study is to present an adaptation to the AnTSbe method, specially designed to be used with fluorescence spectroscopy data. The main idea is to add another feature selection layer to the routine, focused on selecting a Delta Pair (ΔP). The ΔP is one Ex/Em pair that will function as a regulator of the spectra. After been selected, every other valid pair in the spectra will have the ΔP fluorescence intensity subtracted (hence the name delta). As we lose second-order advantages (the property of multi-way calibrations models, as PARAFAC, to have reliable predictions in the presence of un-calibrated constituents (SILVA et al., 2019)) of EEM fluorescence when applying feature selection, the ΔP can act as a reference point to make the fitted models more robust against unknown fluorescence modifiers that were not considered during calibration. Such modifiers could be, for example, unmeasured changes in pH or temperature, the presence of fluorescence quenchers, or contaminants. The use of reference values is a standard practice in optical chemometrics (PAQUET-DURAND et al., 2017). Usually, the user defines

a “blank” spectrum with its intensity reduced from all other spectra. The main objective of removing this blank is to mitigate the environment's influence and avoid spectra variation due to unknown factors (VASAFI et al., 2021). As our proposed methodology is based on feature selection and its advantages, the ΔP has a similar function as the blank spectrum, condensing the reference for the unmeasured changes in the process is this unique pair.

The implementation of the ΔP feature selection will follow the same ACO structure applied in AnTSbe. A new and independent pheromone vector will be created - τΔP. At the beginning of each iteration, each ant will select a ΔP using this pheromone trail and a random trigger. The selected Ex/Em pair is removed from the valid inputs, and its fluorescence intensity is subtracted from other pairs. The selection of model variables and fitting of the models is kept the same. At the end of the iteration, the ant will deposit the same increment in pheromones in the model-selected variables and the selected ΔP. Both pheromone trails are evaporated, and the optimization follows to the next iteration. The selected ΔP will not be part of the tabu memory list. The Global solution will present the best-fitted models, the selected ΔPs, final pheromone trails, and all comparative metrics. Figure 4.1 (B) presents a comparative schematic representation of the proposed adaptation.

An essential step to the implementation of the ΔP is the use of basis expansion. As shown in Equation 4.1, where ŷ is the model predicted output, the introduction of a ΔP in linear models has the same effect as adding another variable. Only with an expansion of two or higher than the ΔP can express significant changes, as shown in Equation 4.2 (model size 2 and σ = 2).

$$\hat{y} = \alpha_1(X_1 - X_{\Delta P}) + \alpha_2(X_2 - X_{\Delta P}) + \beta$$

$$\hat{y} = \alpha_1 X_1 + \alpha_2 X_2 - \left(\sum_1^2 \alpha_i X_{\Delta P}\right) + \beta \tag{4.1}$$

$$\hat{y} = \alpha_1(X_1 - X_{\Delta P}) + \alpha_2(X_2 - X_{\Delta P}) + \alpha_3(X_1 - X_{\Delta P})^2 + \alpha_4(X_2 - X_{\Delta P})^2 + \alpha_5(X_1 - X_{\Delta P})(X_2 - X_{\Delta P}) + \beta \tag{4.2}$$

As the second proposal of this work, the adapted AnTSbe algorithm will be applied to the fluorescence EEM data to optimize feature selection to predict the total phenolic concentration of the aged *cachaças*. The general optimization parameters can be seen in Table 4.1.

Table 4.1. Adapted AnTSbe general optimization parameters.

| **Parameter** | **Value** | **Parameter** | **Value** |
|---|---|---|---|
| OptMetric | RMSE | Basis expansion - σ | 1 and 2 |
| N. Runs - μ | 50 | Model Type | *Ridge* |
| N. iterations - t | 150 | Initial pheromone value – $\tau_0$ | 1000 |
| $N_{ants}$ | 300 | Initial pheromone value – $\tau_{\Delta P}$ | 1000 |
| Tabu Memory Size - z | 5 | Pheromone gain – k | 100 |
| Model Size - Nw | 3 to 6 | Pheromone evaporation rate – ρ | 0.25 |

To quantify the predictive quality of the models, the Root Mean Square Error (RMSE) and Mean Absolute Error (MAPE) metrics were used. They are calculated according to Equations 4.3 and 4.4, respectively, where ŷ is the model predicted output, y is the observed output and *n* number of samples.

$$RMSE = \sqrt{\sum_i^n \frac{(\hat{y}i - yi)^2}{n}} \quad (4.3) \qquad MAPE = \frac{100\%}{n} \sum_i^n \left| \frac{(\hat{y}i - yi)}{\hat{y}i} \right| \quad (4.4)$$

## 4.3 Results and Discussion

### *4.3.1 Quantification of Total Phenolic Concentration*

During the aging processes, the gradual increase in phenolic concentration can be followed in Figure 4.2. The aging process CA1 was faster and more prominent in extracting phenolic compounds from the barrel. On the 7$^{th}$ day of aging, it already had 523.56 mg GAE L$^{-1}$, going up to 1561.75 mg GAE L$^{-1}$ at the end of the experiment (463 days). Aging CA2 and CA3 kept a very similar slower rhythm, assuming a linear increase after 15 days. CA2 reached 577 mg GAE L$^{-1,}$ and CA3 reached a maximum of 497 mg GAE L$^{-1}$, after 204 days. The same barrel was used in aging CA1 and CA2, and we can see by the data that both processes had completely different phenolic profiles. CA3 also used a 20-liter virgin amburana barrel, but its profile was closer to CA2 than CA1. The 'weaker' aging could be correlated to the internal paraffin coating, the hot washes performed, and the smaller percentage of ethanol in the cachaça used in CA2 and CA3. According to Miranda *et al.* (2006) several factors may influence the efficiency of the extraction of wood compounds, such as wood species, barrel age, size, pretreatment of the barrel, environmental conditions, aging time, and alcoholic content of the beverage.

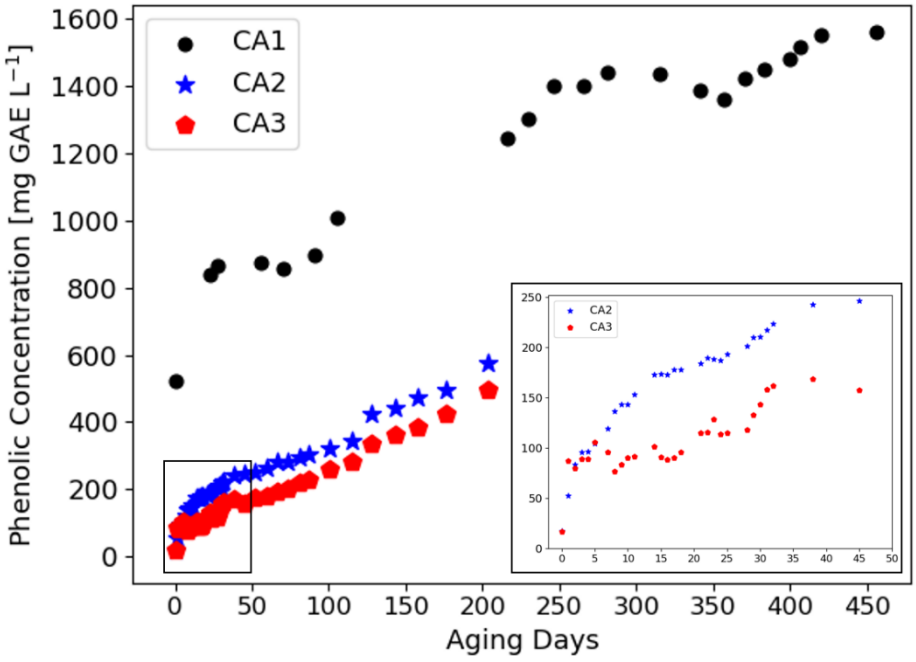



Figure 4.2. The phenolic concentration of aging cachaças by time. CA$_x$ indicates the aging process.

The concentration of total phenolic compounds of the eight commercial *cachaças* evaluated in our research can be seen in Table 4.2. The same distillery produced spirits

Amb02 and Amb03 and, according to the label, were aged by the same time, but Amb03 has twice the concentration of total phenolic compounds, resulting in a lack of product standardization.

Table 4.2. Phenolic concentration of aged commercial cachaças.

| Name | Phenolic Conc. [mg GAE $L^{-1}$] | Aging [years] | Name | Phenolic Conc. [mg GAE $L^{-1}$] | Aging [years] |
|---|---|---|---|---|---|
| Amb01 | 78.56 | 1 | Mix01 | 50.21 | 6 |
| Amb02 | 132.16 | 1 | Blend01 | 57.85 | 2 |
| Amb03 | 264.36 | 1 | Oak01 | 116.47 | - - - |
| Balm01 | 82.82 | 1 | Cab01 | 166.46 | - - - |

Related works such as Santiago *et al.* (2017) quantified the phenolic concentration of five cachaças aged in Jatobá wood for one year, finding concentrations between 281.85 and 876.98 mg GAE $L^{-1}$. Bernardes *et al.* (2014) quantified phenolic concentration in 103 samples of Brazilian aged cachaças, finding values ranging from 0.73 to 82.34 mg GAE $L^{-1}$. As phenolic compounds can be directly correlated to sensorial properties of the spirit (color, odor, and flavor), this massive difference between aging processes is an indication that a more practical phenolic quantification method can be very beneficial to the industry. This information can be used to define appropriate aging times, make products more consistent year-round, and even determine the lifespan of aging barrels.

### *4.3.2 EEM Fluorescence Spectroscopy*

The fluorescence spectra were measured without any sample pretreatment. Figure 4.3 shows the spectrum of eight cachaças going from lower to higher phenolic content. The red ploted contour represent the maximum fluorescence intesity in each excitation wavelength, and the blue contour the maximum fluorescence for each emission wavelength. In these spectra, it is possible to observe significant changes in the regions of fluorescence. There is a tendency in both CA2 and CA3 of the displacement of the fluorescence peak to the right-hand of the spectra. The same tendency can be seen in the work of Carvalho *et al.* (2020), which first studied the CA1 aging and constructed laboratory-made samples to simulate aged cachaças with 0 – 450 mg GAE $L^{-1}$. This work confirms that natural aging processes also present the fluorescence peak shift observed in diluted samples. For further information, the Appendix presents a collection of EEM spectra of each CA and commercial cachaças.

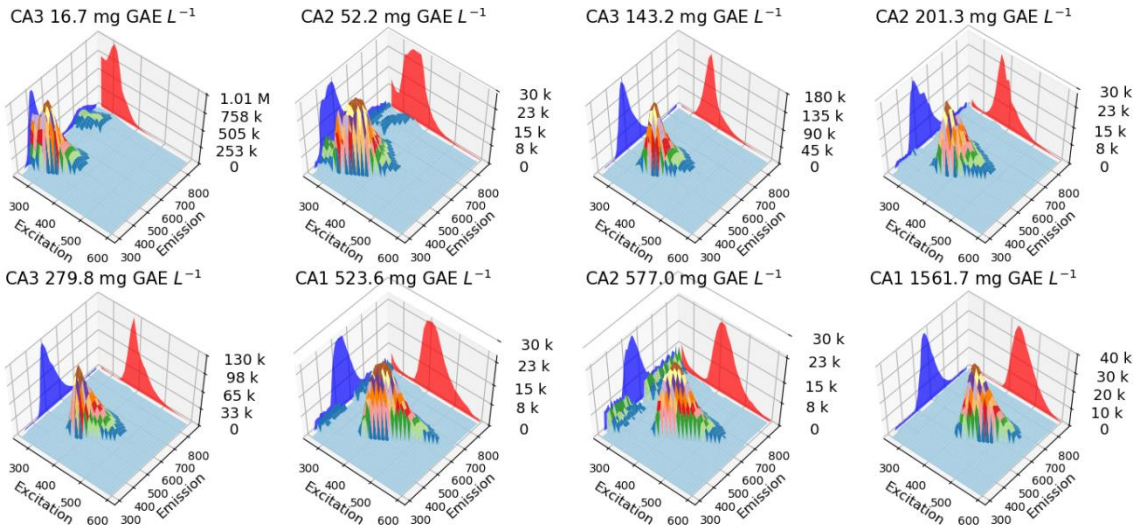



Figure 4.3. EEM fluorescence spectra of cachaças with increasing phenolic concentration. Above each subplot, the $CA_x$ refers to the aging process, followed by the sample's total phenolic content.

Another piece of information drawn from the individual EEM is the vast difference in fluorescence intensity. It is shown in Figure 4.4, where the mean fluorescence intensity of each spectrum is plotted against that sample's total phenolic content. Samples with smaller phenolic concentrations had, in general, higher mean fluorescence intensity. The mean intensity gradually decreases until it reaches a plateau and then slowly rises again. The decreasing behavior can be seen following CA2 and CA3. The average intensity of CA1 samples, compared to the other CAs, has hardly changed, even though the total phenolic content has tripled. Processes CA1 and CA2 (that used the same barrel but different cachaças) have similar averages where their phenolic content intersects, looking like they are just one continuous process. On the other hand, CA3 samples had an average of up to 10 times higher than CA2, having a profile closer to the commercial spirits. The individual plots can be seen in the Appendix.

As the data suggests, the changes in spectra can be associated with the different compounds extracted from the barrel during aging. Volatile oils, tannins, phenols, non-volatile organic acids, and sugars are the main compounds extracted from wood (CARDELLO; FARIA, 2000). A variety of molecular interactions can cause a decrease in fluorescence intensity, as internal absorption, excited-state reactions, molecular rearrangements, energy transfer, and collisions (LAKOWICZ, 2006). Also, highly fluorescent components may suppress emission from components with low quantum yields resulting in the disappearance of bands (SIKORSKA; KHMELINSKII; SIKORSKI, 2019).

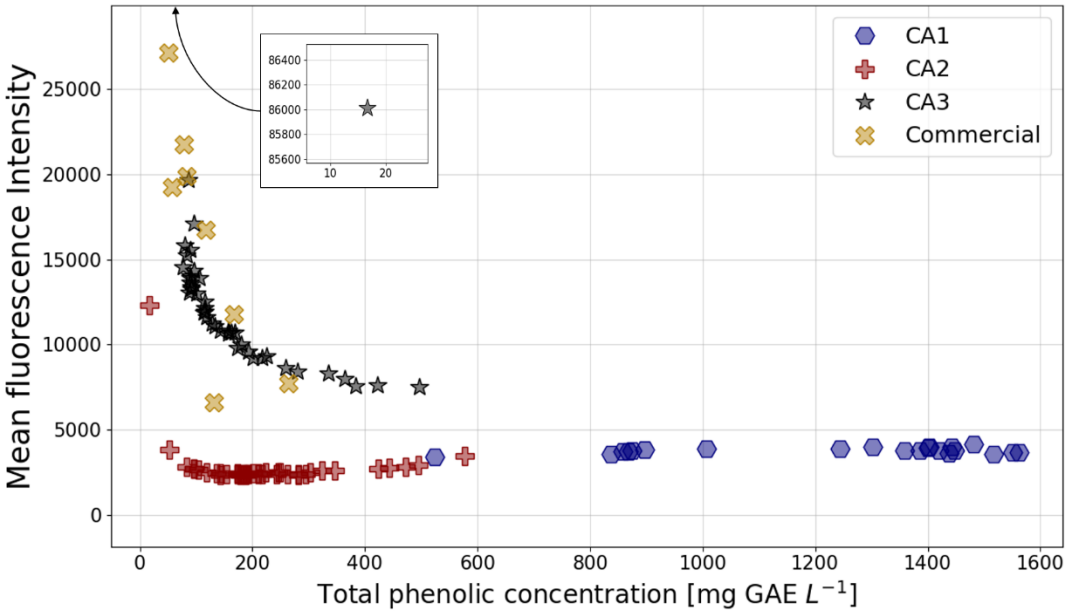



Figure 4.4. Mean EEM fluorescence intensity *vs* total phenolic concentration of each sample.

### *4.3.3 Chemometric analysis – Adapted AnTSbe algorithm*

In the first analyses, all available samples were used, with the restriction that the eight commercial cachaças were allocated only to the validation and test subsets (four each). This way, we can evaluate the ability of the models to deal with samples that differ significantly from the calibration subset.

Table 4.3. Metrics, selected fluorescence pairs, and ΔP for the MD predictive models with model size 5, using all available samples.

| | RMSE (mg GAE $L^{-1}$) | | | MAPE ( % ) | | |
|---|---|---|---|---|---|---|
| Parameters | Cal. | Val. | Test | Cal. | Val. | Test |
| **MD01** σ = 1, ΔP = False | 58.20 | 73.70 | 69.50 | 32.4 | 31.9 | 38.0 |
| **MD02** σ = 2, ΔP = False | 35.80 | 49.50 | 74.60 | 10.2 | 17.4 | 39.7 |
| **MD03** σ = 2, ΔP = True | 33.90 | 52.00 | 43.90 | 13.3 | 23.0 | 18.9 |
| **Selected Excitation/Emission Pairs** | | | | | | |
| **MD01** | Ex400/Em700 | Ex410/Em670 | Ex440/Em520 | Ex480/Em560 | Ex540/Em840 | |
| **MD02** | Ex280/Em760 | Ex380/Em530 | Ex440/Em520 | Ex460/Em550 | Ex480/Em580 | |
| **MD03** | Ex270/Em500 | Ex360/Em660 | Ex360/Em670 | Ex430/Em510 | Ex580/Em820 | |
| | **ΔP - Ex470/Em560** | | | | | |

Three arrangements were tested: basis expansion of 1 and no ΔP – MD01, basis expansion of 2 and no ΔP – MD02, and basis expansion of 2 and ΔP – MD03. For each arrangement, the adapted AnTSbe is run using the parameters defined in Table 1. Models' sizes from 3 to 6 were tested. Table 4.3 presents the metrics, selected fluorescence pairs,

and ΔPs for the predictive models based on EEM fluorescence, using all available samples. In all three arrangements, there was no significant improvement in metrics going from model size 5 to 6, so model size 5 was presented.

For a more visual comprehension, Figure 4.5 shows the predicted *versus* measured outputs for the MD fitted models. Evaluating the metrics, we can see that the use of basis expansion was beneficial for the predictive models. The introduction of higher-order variables, on the other hand, also exacerbates deviations from the standards learned from calibration samples. It can be seen by the MD02 inability to correctly predict commercial test samples, although being better than MD01 in every other aspect (including commercial validation and regular test samples).

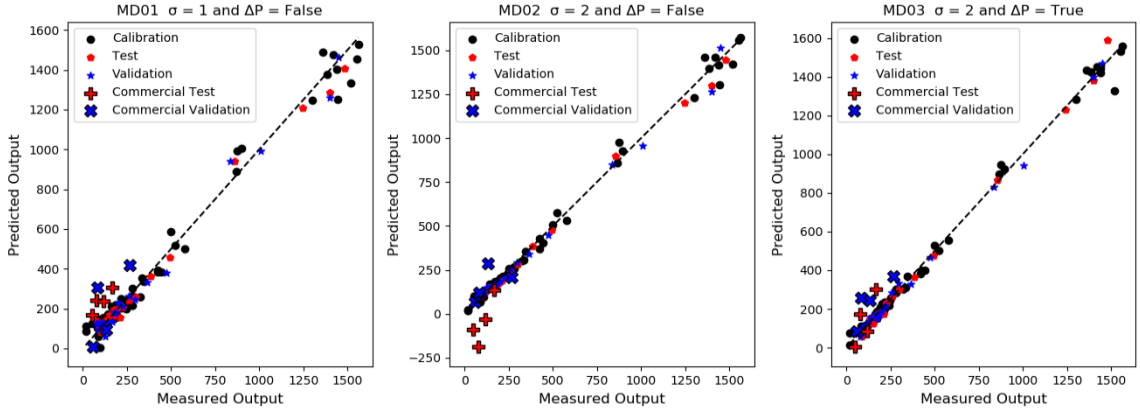



Figure 4.5. Predicted *vs.* Measured outputs for the AD arrangements, using all available samples and model size 5.

Comparing MD02 and MD03, the use of ΔP enhanced the generalization potential of the model, softening the expansion's adverse effects on commercial test samples.

Contemplating the significant difference in concentration between CA2/CA3/Commercial samples and CA1, and the apparent shift of fluorescence bands aforementioned, the study of local models, focused on a smaller phenolic range, can be beneficial.

To fit the new local models, only aging CA2, CA3, and Amburana commercial samples (Amb01, Amb02, and Amb03 – incorporated into the test subset) were considered. Parameters, model sizes, and arrangements were kept the same (renamed to AD0x).

Table 4.4 presents the metrics, selected fluorescence pairs, and ΔP for the predictive AD models. As before, there was no significant improvement between model sizes 5 to 6, so model size 5 metrics are displayed.

Table 4.4. Metrics, selected fluorescence pairs, and ΔP for the predictive AD models with model size 5, using only CA2, CA3 and Amburana commercial samples.

| | RMSE (mg GAE $L^{-1}$) | | | MAPE ( % ) | | |
|---|---|---|---|---|---|---|
| Parameters | Cal. | Val. | Test | Cal. | Val. | Test |
| **AD01** σ = 1, ΔP = False | 26.30 | 18.20 | 26.00 | 21.60 | 10.80 | 13.10 |
| **AD02** σ = 2, ΔP = False | 11.00 | 4.00 | 89.90 | 8.10 | 2.20 | 36.20 |
| **AD03** σ = 2, ΔP = True | 11.90 | 8.50 | 28.00 | 8.70 | 3.70 | 13.20 |
| **Selected Excitation/Emission Pairs** | | | | | | |
| **AD01** | Ex420/Em520 | Ex440/Em520 | Ex480/Em560 | Ex490/Em590 | Ex490/Em680 | |
| **AD02** | Ex340/Em520 | Ex370/Em550 | Ex440/Em670 | Ex490/Em590 | Ex500/Em650 | |
| **AD03** | Ex290/Em460 | Ex360/Em650 | Ex380/Em460 | Ex440/Em520 | Ex510/Em830 | |
| | **ΔP - Ex480/Em560** | | | | | |

Figure 4.6 complements the AD model metrics, showing the predicted *vs.* measured outputs. The results follow the same trend of the MD: the use of basis expansion improved general model predictions but made the AD02 model fail more on the commercial samples. The introduction of the ΔP once again was crucial to the generalization capability of the model. Not considering the commercial test samples, both AD02 and AD03 models had similar RMSE and MAPE metrics for the test subset: around 13 mg GAE $L^{-1}$ and 6.5%, respectively. In this case, both the original AnTSBe (AD02) and the proposed Adapted AnTSBe (AD03) could capture the process behavior aptly. The AD03 model was not only able to satisfactorily predict total phenolic concentration in two distinct aging processes, as was capable of reasonably predicting the concentration in commercial samples, not involved in the calibration procedures.

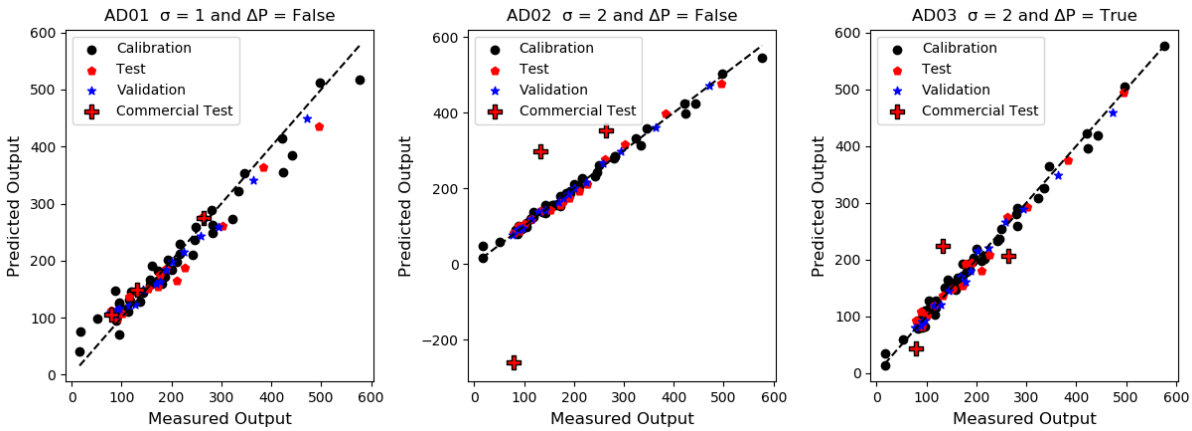



Figure 4.6. Predicted vs. Measured outputs for the AD arrangements, using only CA2/CA3 and Amburana commercial samples, with model size 5.

Although entirely independent from each other, both MD03 and AD03 selected similar Ex/Em pairs and ΔPs. Figure 4.7 shows the pairs' locations plotted upon the average EEM spectra of all samples. The selection of similar pairs is beneficial for the feature selection procedure, indicating that less information needs to be acquired, and only the model

parameters must be adjusted, when dealing with local models within specific a phenolic range.

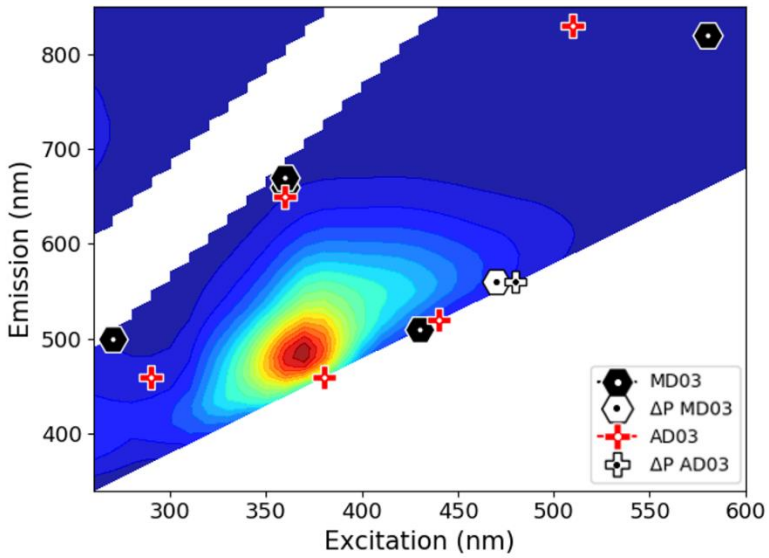



Figure 4.7. MD03 and AD03 selected fluorescence Ex/Em pairs and ΔPs (model size 5).

## 4.4 Conclusions

This paper proposed adapting the stochastic feature selection algorithm AnTSbe to be combined with EEM fluorescence spectra for quantification purposes. The methodology consists of adding an external feature-selecting layer to optimize the selection of a ΔP that will function as a regulator of the spectra. The adapted algorithm was applied to cachaça's aging fluorescence data to quantify total phenolic concentration. Data from three lab-aging processes and eight commercial, aged cachaças were used. Although similar in general characteristics, each aging process had very different phenolic and fluorescence profiles over time. Besides, the evaluation of the EEM spectra shows the tendency of the fluorescence peak to shift to the right-hand of the spectra with the increase of phenolic content, and the suppression of fluorescence bands.

The obtained results showed that the polynomial combination of input features improved the predictive capability of the MD models. However, the model was more sensible to distinguished data and was unable to predict commercial test samples satisfactorily. The introduction of the ΔP softened the influence of higher-order variables and improved the generalization capability of the models. Using only 5 out of the original 1065 input variables in a global model with all samples, we could predict total phenolic content adequately, considering the high amplitude of the output. The use of the ΔP reduced the test RMSE and MAPE from 74.60 mg EAG $L^{-1}$ and 39.7 % (MD02) to 43.9 mg EAG $L^{-1}$ and 18.9 % (MD03).

When fitting local models using only CA2, CA3, and commercial cachaças aged in Amburana, the same trend was observed. The use of polynomial combinations improved the model generally but made it worst when predicting unfamiliar commercial spirits. ΔP was crucial to reduce test MAPE from 36.2% (AD02) to 13% (AD03). When no commercial samples were considered, both AD02 and AD03 achieved similar, satisfactory results, with MAPE around 6.5%. Although independent and using unmatched data, both MD03 and AD03 models selected similar Ex/Em pairs and ΔPs. The results shown corroborate the

proposed methodology and support that the adapted AnTSbe was necessary to improve the model's generalization capability.

## 4.5 Acknowledgments

The authors are grateful to Cachaçaria Weber Haus and thankful to CAPES for financial support.

## 4.6 References

ASSAWAJARUWAN, S.; REINALTER, J.; HITZMANN, B. Comparison of methods for wavelength combination selection from multi-wavelength fluorescence spectra for on-line monitoring of yeast cultivations. **Analytical and Bioanalytical Chemistry**, v. 409, n. 3, p. 707–717, 2017.

BAHRAM, M. et al. Handling of Rayleigh and Raman scatter for PARAFAC modeling of fluorescence data using interpolation. **Journal of Chemometrics**, v. 20, n. 3–4, p. 99–105, 1 mar. 2006.

BECKER, T. et al. Future Aspects of Bioprocess Monitoring. In: ULBER, R.; SELL, D. (Eds.). . **White Biotechnology**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 249–293.

BERNARDES, C. D.; DE FIGUEIREDO, M. C. P.; BARBEIRA, P. J. S. Developing a PLS model for determination of total phenolic content in aged cachaças. **Microchemical Journal**, v. 116, p. 173–177, 2014.

CARDELLO, H. M. A. B.; FARIA, J. B. Análise da aceitação de aguardentes de cana por testes afetivos e mapa de preferência interno. **Ciênc. Tecnol. Aliment.**, v. 20(1), p. 32- 36., 2000.

CARVALHO, D. G. et al. Determination of the concentration of total phenolic compounds in aged cachaça using two-dimensional fluorescence and mid-infrared spectroscopy. **Food Chemistry**, v. 329, p. 127142, 1 nov. 2020.

FAASSEN, S. M.; HITZMANN, B. Fluorescence Spectroscopy and Chemometric Modeling for Bioprocess Monitoring. p. 10271–10291, 2015.

IBRAC. **Instituto Brasileiro de Cachaça. Brasil. Mercado interno e mercado externo.**, 2019.

ITAKURA, K. et al. Estimation of Citrus Maturity with Florescence Spectroscopy Using Deep Learning. **Horticulturae**, v. 5, n. 1, p. 2, 26 dez. 2018.

KETTANEH, N.; BERGLUND, A.; WOLD, S. PCA and PLS with very large data sets. **Computational Statistics & Data Analysis**, v. 48, n. 1, p. 69–85, 2005.

KUMAR, N. et al. **Chemometrics tools used in analytical chemistry: An overviewTalanta**Elsevier B.V., , 1 jun. 2014.

LAKOWICZ, J. R. **Principles of Fluorescence Spectroscopy**. 3. ed. Baltimore, Maryland, USA: [s.n.].

MIRANDA, M. B. DE; HORII, J.; ALCARDE, A. R. Estudo Do Efeito Da Irradiação Gamma ( 60 Co ) Na Qualidade. v. 26, n. 4, p. 772–778, 2006.

MOSEDALE, J. R.; PUECH, J. L. **Wood maturation of distilled beveragesTrends in Food Science and Technology**Elsevier Sci Ltd, , 1 mar. 1998.

ÖDMAN, P. et al. On-line estimation of biomass, glucose and ethanol in Saccharomyces cerevisiae cultivations using in-situ multi-wavelength fluorescence and software sensors. **Journal of Biotechnology**, v. 144, n. 2, p. 102–112, 2009.

PAQUET-DURAND, O. et al. Artificial neural network for bioprocess monitoring based on fluorescence measurements: Training without offline measurements. **Engineering in Life Sciences**, v. 17, n. 8, p. 874–880, 9 out. 2017.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, n. Oct, p. 2825–2830, 2011.

RANZAN, L.; TRIERWEILER, L. F.; TRIERWEILER, J. O. Prediction of sulfur content in diesel fuel using fluorescence spectroscopy and a hybrid ant colony - Tabu Search algorithm with polynomial basis expansion. **Chemometrics and Intelligent Laboratory Systems**, v. 206, p. 104161, 15 nov. 2020.

SANTIAGO, W. D.; CARDOSO, M. DAS G.; NELSON, D. L. Cachaça stored in casks newly constructed of oak (Quercus sp.), amburana (Amburana cearensis), jatoba (Hymenaeae carbouril), balsam (Myroxylon peruiferum) and peroba (Paratecoma peroba): alcohol content, phenol composition, colour intensity and dry extrac. **Journal of the Institute of Brewing**, v. 123, n. 2, p. 232–241, 2017.

SANTOS, P. V. J. L. et al. K-RANK: AN EVOLUTION OF Y-RANK FOR MULTIPLE SOLUTIONS PROBLEM. **Brazilian Journal of Chemical Engineering**, v. 36, p. 409–419, 2019.

SIKORSKA, E.; KHMELINSKII, I.; SIKORSKI, M. Fluorescence spectroscopy and imaging instruments for food quality evaluation. In: **Evaluation Technologies for Food Quality**. [s.l.] Elsevier, 2019. p. 491–533.

SILVA, A. C. et al. Green chemistry method based on PARAFAC EEM data modeling for Benzo[a]pyrene quantitation in distilled spirit. **Journal of the Brazilian Chemical Society**, v. 30, n. 2, p. 398–405, 2019.

SINGLETON, V. L.; ROSSI, J. A.; JR, J. COLORIMETRY OF TOTAL PHENOLICS W I T H A C I D REAGENTS. 1985.

VASAFI, P. S. et al. Anomaly detection during milk processing by autoencoder neural network based on near-infrared spectroscopy. **Journal of Food Engineering**, v. 299, p. 110510, 1 jun. 2021.

## 4.7 Appendix A

Figures A1 to A4 present the EEM fluorescence spectra of samples from the CA processes and all commercial cachaças.

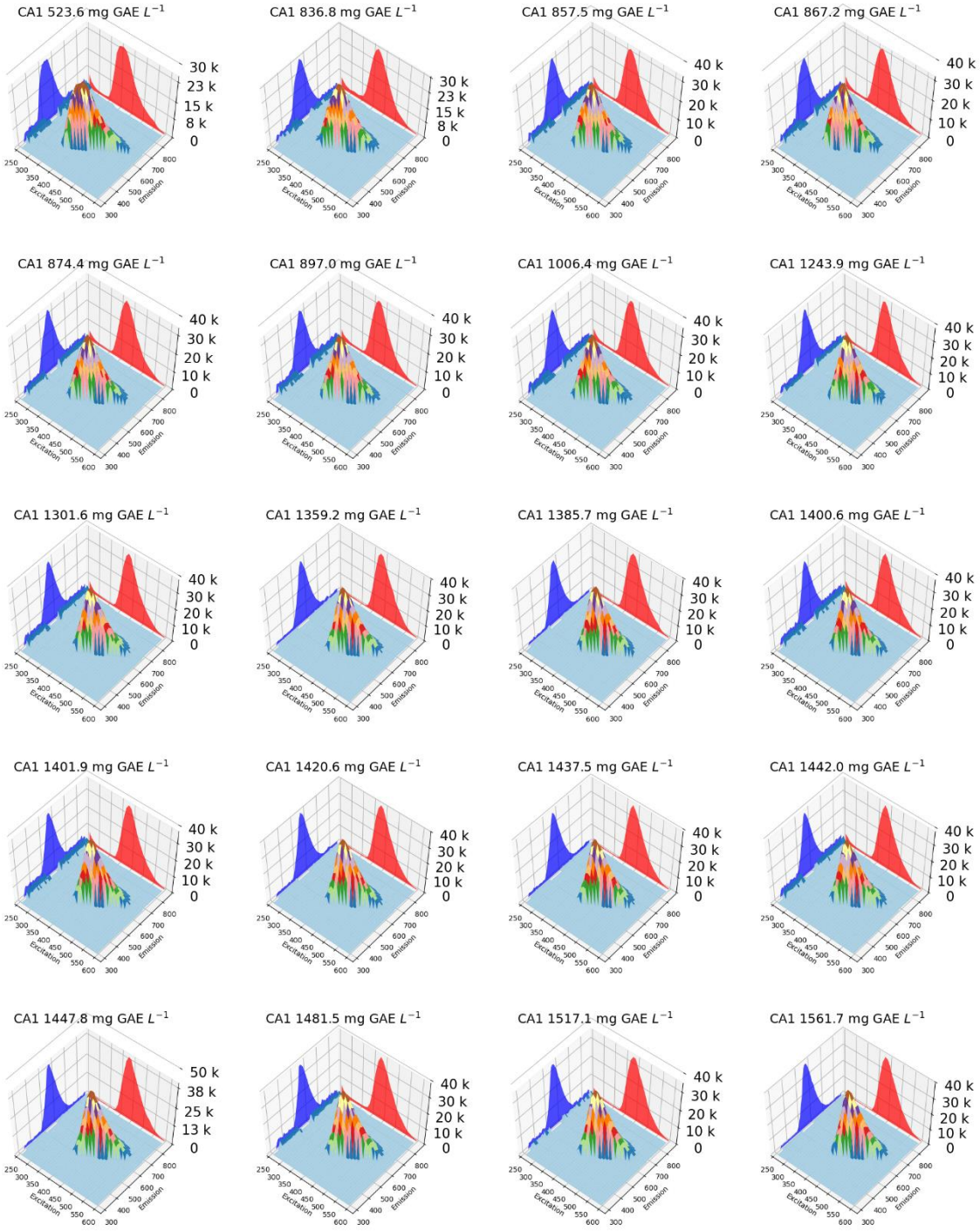



Figure A1. EEM fluorescence spectra of 20 out of 21 samples from CA1 aging process, including first and last collected samples. Above each subplot, the phenolic concentration of that sample is shown.

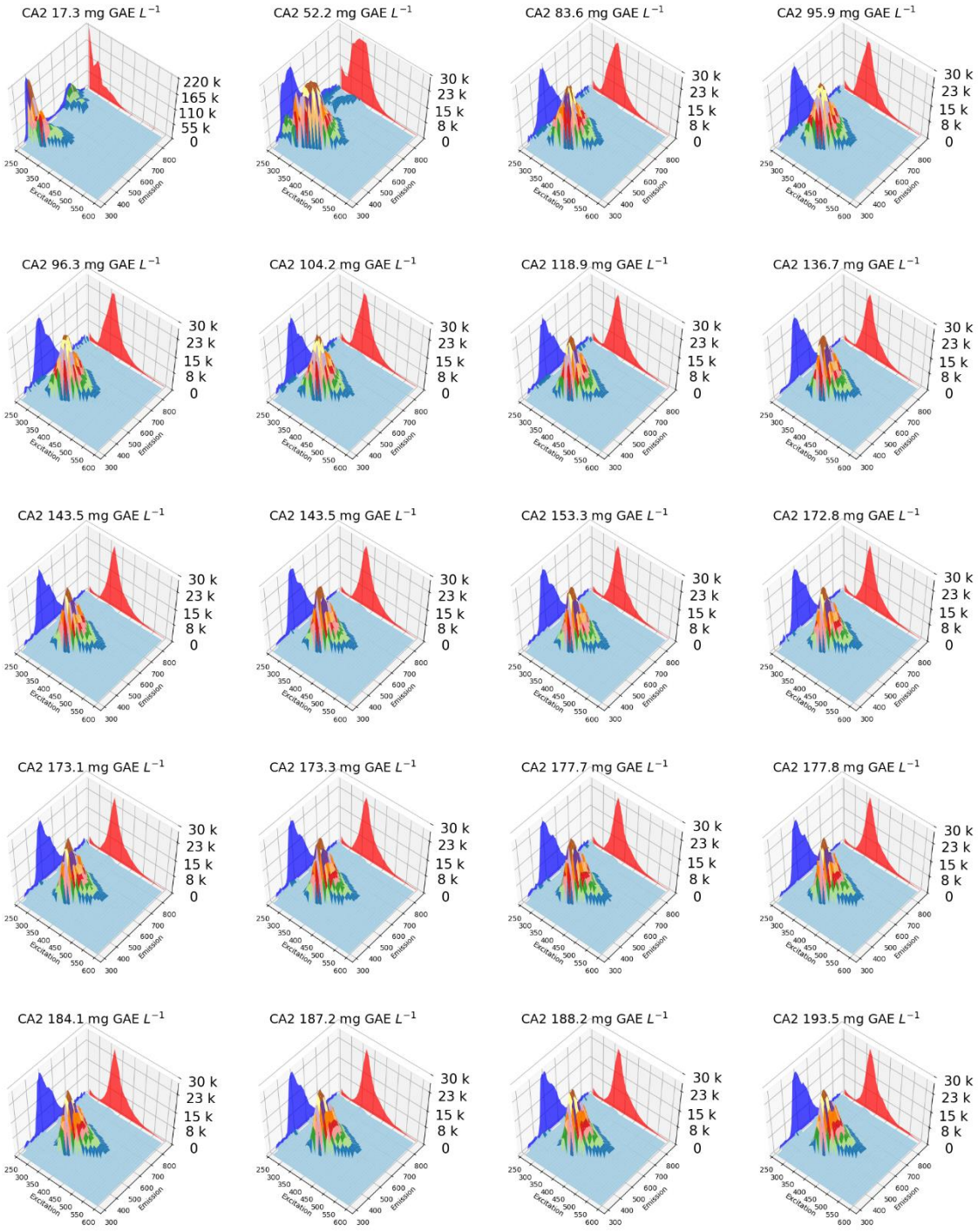



Figure A2. EEM fluorescence spectra of 20 out of 41 samples from CA2 aging process, including first and last collected samples. Above each subplot, the phenolic concentration of that sample is shown.

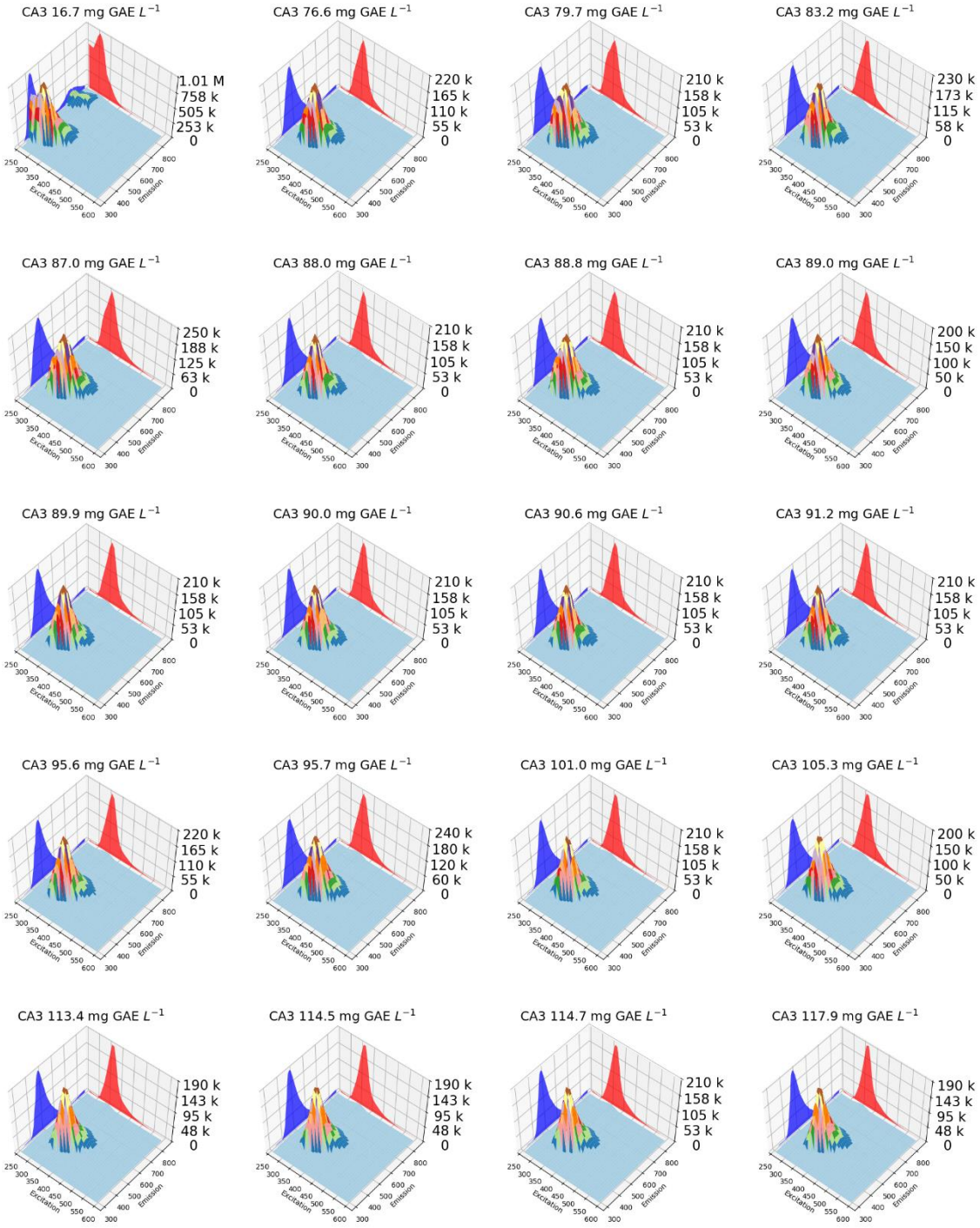



Figure A3. EEM fluorescence spectra of 20 out of 41 samples from CA3 aging process, including first and last collected samples. Above each subplot, the phenolic concentration of that sample is shown.

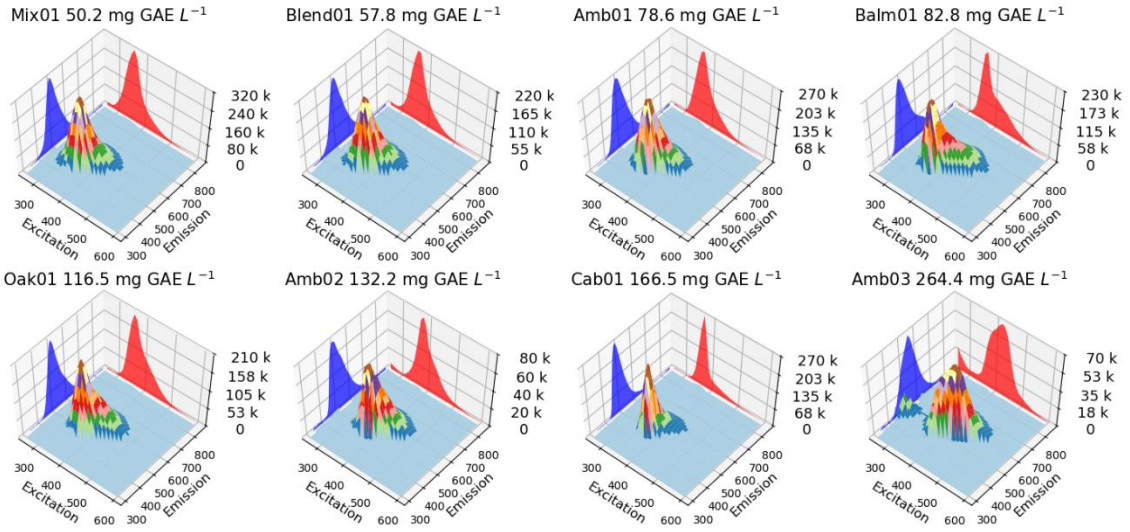



Figure A4. EEM fluorescence spectra of all commercial samples. Above each subplot, the name and phenolic concentration of that sample is shown.

Figure A5 exhibits the mean fluorescence intensity of all samples against their total phenolic concentration.

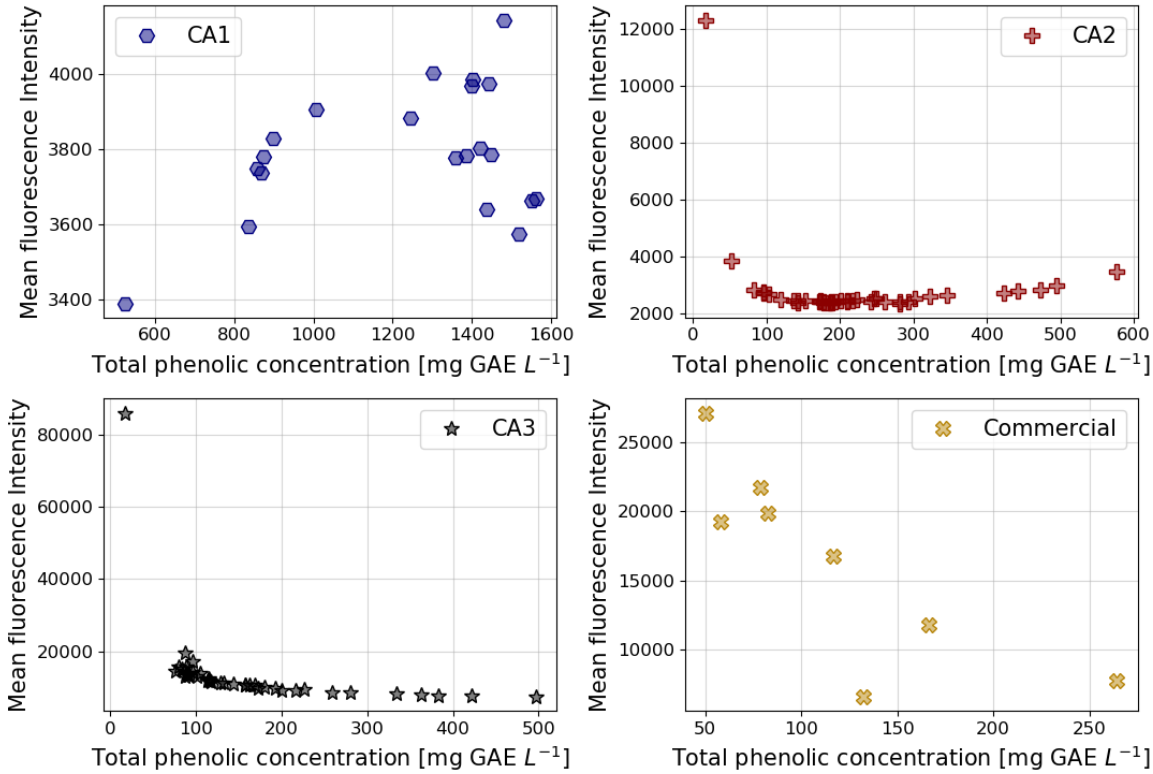



Figure A5. Mean fluorescence intensity of samples *vs.* their total phenolic concentration. $CA_x$ indication refers to the aging processes.

# Capítulo 5 – **Avoiding Misleading Predictions in Fluorescence-based Soft Sensors using Autoencoders**

**Abstract:** Fluorescence spectroscopy is a powerful tool for monitoring fermentation processes in autofluorescence microorganisms like yeast. The characteristics of the excitation-emission matrix fluorescence data can be explored by using deep convolutional neural networks to enhance the predictive performance of bioprocess variables. This article proposes the use of residual neural networks (ResNet) for the prediction of ethanol, glucose, and biomass concentrations of *S. cerevisiae* cultivations based on fluorescence data collected *in situ*. In addition, to address the problem of only ensuring the quality of future predictions in data similar to the training dataset, we propose a trust screening based on autoencoder (AE) reconstruction error. Its characteristic of reconstructing the inputs is the key feature to avoid misleading predictions and forecast if a new sample should be trusted as usual or flagged as abnormal. The 83 layers deep ResNet successfully predicted the desired outputs, with $R^2$ higher than 0.99 and Root Mean Square Error (RMSE) of 0.02325, 0.02410, and 0.03717 g/L for the biomass, ethanol, and glucose concentrations, respectively, in the test subset (0.72%, 0.75% and 0.35%, relative to the range of the extrapolated offline data). The best-fitted autoencoder had a 3-layer architecture, with three neurons in the bottleneck and using rectified linear unit (ReLU) activation for the encoder and linear activation for the decoder. The RMSE for the fermentations was 4.91 rel. Fluorescence intensity units, representing a general error smaller than 0.65%. To evaluate the AE capability to work as trust screening, random fluorescence intensity was added to the Ex450/Em530 fluorescence pair (related to flavins) in some samples, creating a defective dataset. The dataset was evaluated with the trained AE and the ResNet model to compare reconstruction errors and bioprocess concentrations. The AE was able to identify the samples with added errors, and, as expected, the defective samples also presented higher predictive errors in general. The higher the reconstruction RMSE, the less the new sample should be trusted to avoid misleading predictions.

## 5.1 Introduction

Two-dimensional fluorescence spectroscopy has been a valuable tool for monitoring cultivation processes, especially for autofluorescence microorganisms like yeast (MASLANKA; KWOLEK-MIREK; ZADRAG-TECZA, 2018; FAASSEN; HITZMANN, 2015). Fluorescence spectroscopy is a highly efficient optical method that can be used to simultaneously measure many constituents of living cells (PODRAZKÝ et al., 2003). Also, non-invasive optical techniques are ideal for dealing with biological systems because it does not interfere with microbes or cells inside a bioreactor. A fair number of articles were published presenting different approaches for using fluorescence spectra in soft sensors. Podrazky *et al.* (2003) used two-dimensional fluorescence to monitor growth and stress responses to different treatments in *S. cerevisiae* cultivations. Rhee and Kang (2007) examined principal component regression and partial least squares to find a correlation between fluorescence spectroscopy and bioprocess variables of recombinant *Escherichia coli* cultivations. Masiero *et al.* (2013) evaluated different methods of wavelength selection for the characterization of bioprocesses, comparing Exhaustive Search (ES), Stepwise Regression, and Genetic Algorithm (GA) for the selection of few excitation/emission fluorescence pairs that could be used in multi-linear chemometric models to predict biomass, ethanol and glucose concentration in *S. cerevisiae* fermentations. For the problem stated in their work, only Stepwise Regression presented unsatisfactory performance, with ES and GA achieving $R^2$ higher than 0.98 for the three bioproducts. In a similar study, Assawajaruwan *et al.* (2017) compared wavelength selection methods for the online monitoring of *S. cerevisiae* cultivations. They compared methods based on principal component loadings, variable importance in projection, and ant colony optimization to select relevant wavelength combinations to predict glucose, ethanol, and biomass concentration. All three methods performed well, selecting fluorescence pairs in the same regions of biogenic fluorophores related to yeast metabolism, as NAD(P)H, tryptophan, pyridoxine, and riboflavin.

As a nonlinear alternative to all previously cited methodologies, artificial neural networks gained massive expression in later years (CHARTE et al., 2018). Paquet-Durand *et al.* (2017) proposed using a feed-forward neural network combined with fluorescence spectroscopy to monitor *S. cerevisiae* fermentation where no offline measurement was needed. A theoretical model of the process is applied to simulate biomass, glucose, and ethanol concentration. The kinetic parameters of the simulation model and the parameters of the feed-forward neural network are acquired from the 2D fluorescence spectra alone. The resulting trained neural network predicted the process state accurately as the conventionally (with offline measurements) trained neural network.

In all approaches so far, the fluorescence spectra are treated as a combination of individual excitation/emission pairs; each pair has been considered a standalone variable, not taking advantage of the multi-dimensional characteristic of the data. Itakura *et al.* (2018) evaluated citrus maturity by predicting the Brix/acid ratio of juice from the flesh of mandarins, using excitation-emission matrix (EEM) fluorescence spectroscopy and deep learning. In their study, the EEM is regarded as an image, and the maturity is estimated via performing a regression with Convolutional Neural Network (CNN regression). The authors reported an absolute Brix/acid ratio error considerably better than the values obtained by previous studies. Rutherford *et al.* (2020) also applied EEM fluorescence spectroscopy and CNN to predict the source of combustion-generated particulate. The authors compiled the EEM spectra of smoke from cigarettes, diesel, and wood and used a CNN to identify the

presence or absence of known particular matter sources in the EEM spectra. The CNN was able to identify cigarette and woodsmoke individually and in the presence of other sources with 98% and 99% accuracy, respectively. The overall classification accuracy for all three sources was 89%.

The use of deep CNN led to a series of breakthroughs, especially in image classification (ZEILER; FERGUS, 2014). With constant increments in layers and deeper networks able to start converging, a degradation problem was revealed: the increment in-depth saturated the accuracy, and it was not by overfitting. In theory, adding more layers to a suitable model should not produce higher training errors. At maximum, the new layers would just perform identity mapping and copy previous results, but experiments showed detriment in training error caused by the inability of solvers to find this identity solution (HE; SUN, 2014). The introduction of 'shortcut connections' in a deep residual learning framework can address the degradation problem. Figure 5.1 represents the standard “plain” network (left) and a shortcut connection (right), where one or more layers are skipped. The identity shortcut does not add extra parameters or computational complexity to the network, and it is easier to optimize the residual mapping by backpropagation than the original (HE et al., 2016a).

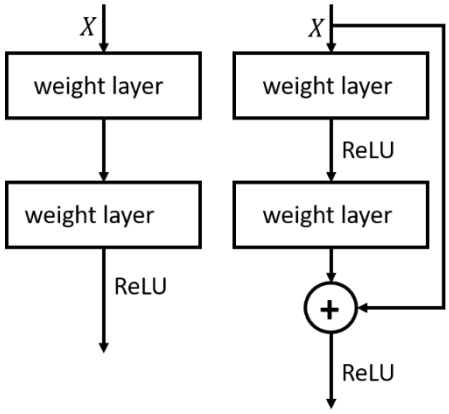



Figure 5.1. Plain network (left) and shortcut connection (right).

The use of this residual learning approach culminated in the renowned residual networks (*ResNets*), deep stacks of tens, or hundreds of CNN, with a shortcut connection added after a set number of filters. For example, the 152-layer *ResNet* proposed by He *et al.* (2016) won the *ILSVRC 2015 classification competition*, being the deepest network presented so far. Very deep residual networks and the evolution in their architecture are effervescent topics and state-of-the-art for several applications (HANIF; BILAL, 2020; HU et al., 2018; IOFFE; SZEGEDY, 2015; LATHUILIÈRE et al., 2018; XUE et al., 2016).

Considering the multi-dimensionality of the EEM fluorescence data and how well fluorescence has been historically used to predict fermentation bioproducts, combining deep residual networks can be a valuable tool to track and control continuous or batch bioprocesses in a non-invasive way.

The approach of using only data and no phenomenological basis to fit models are called “black-box” modeling and is the primary methodology used today to deal with big data and to process digital information (SILVEIRA; COELHO; SANTOS, 2015). The main problem with black-box models is that you can only trust their predictions if the input data is within the

same range of the data used to fit it. Moreover, if faulty or defective data is passed to the model, it can result in misleading predictions. Even for humans, it is difficult to distinguish between normal and abnormal states only by looking at raw data, especially when there are hundreds or thousands of variables. For this reason, it is valuable to train the machine to learn how to recognize differences between common and uncommon states. Sakurada and Yairi (2014) compared the performance of linear PCA, kernel PCA, autoencoders, and denoising autoencoders in the task of anomaly detection. The work analyzed simulated data from the Lorenz system and real data from two types of spacecraft telemetry. The authors concluded that linear PCA fails to identify anomalies in high latent dimensions. The autoencoder and denoising autoencoder performed the same or better than the kernel PCA, which requires more massive computation. In autoencoders, the original and the reconstructed data are in the same original observation space, not needing to solve the complex pre-image problem that kernel PCA requires. Amarbayasgalan *et al.* (2018) also evaluated unsupervised novelty detection using deep autoencoders. The authors compared twelve novelty detection methods from the literature, eleven of which were based on support vector machines or PCA/Kernel PCA. Their proposed method was based on deep autoencoders reconstruction errors and density-based clustering algorithm. For twenty outlier benchmark datasets from the Outlier Detection Datasets (ODDS), the authors conclude that the average AUC of deep autoencoder algorithms was at least 13.5% better than the contestants.

The autoencoder (AE) is the most basic and straightforward architecture in unsupervised deep learning. It consists of an artificial neural network that produces a hidden latent space, i.e., encodes the given data, so that its decodification seeks to replicate the inputs (CHARTE et al., 2018). The schema in Figure 5.2 presents an example of the architecture of an AE. The middle hidden layer, or latent space, can be called a bottleneck due to its format. This symmetric network is trained by pursuing the minimal reconstruction error between the output and input data (which are the same) (JES et al., 2019).

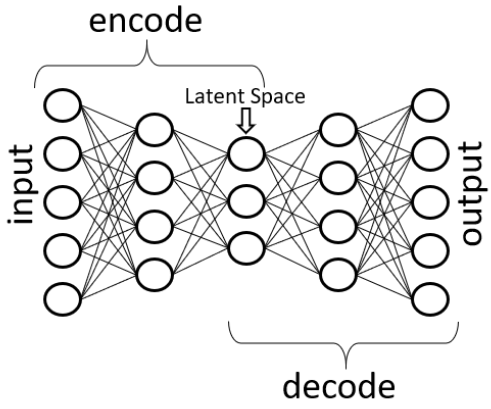



Figure 5.2. Autoencoder: Example schema of the architecture of the neural network.

Similar to Principal Component Analysis – PCA (MANNING-DAHAN, 2017), it is possible to reduce the dimensionality of the original data and compress it into a new space without losing its essence in a lower dimension. The most highlighted difference between PCA and AE is capturing the nonlinear characteristics and the effectiveness of representation in the latent space by the AE (JES et al., 2019). This latent space can be used for visualization purposes or to predict the characteristics of the data. The vast majority of AE applications

are focused on sound (DENG et al., 2014; YAMSHCHIKOV; TIKHONOV, 2017), image and video data (ALMOTIRI; ELLEITHY; ELLEITHY, 2017; DOERSCH, 2016).

The contributions of this work are two-fold. First, we will estimate biomass, ethanol, and glucose concentration in *S. cerevisiae* fermentations via performing a regression with Residual Networks (ResNet) and EEM fluorescence spectroscopy data collected *in situ* in a bioreactor. Second, we propose a trust screening based on autoencoder reconstruction error to validate the EEM spectra, detect abnormalities and avoid misleading predictions. Both methods run in parallel: from the same input data, the ResNet predicts the bioprocess states, and the AE reconstruction error indicates if the data should be trusted.

## 5.2 Materials and Methods

### *5.2.1 Saccharomyces cerevisiae Fermentation Monitoring using EX/EM Fluorescence*

The dataset used in this work was provided by Professor Dr. Bernd Hitzmann, director of the Process Analytics and Cereal Science department of Universität Hohenheim, and presented in the work of Paquet-Durand *et al.* (2017): three batch fermentation of baker's yeast (*S. cerevisiae*), named P1, P2, and P3, were performed. During the fermentation, the two-dimensional fluorescence spectra of the medium were collected *in situ* using a BioView fluorescence spectrometer (DELTA Lights & Optics, Hørsholm, Denmark) in the range of 270–550 nm excitation and 310–590 nm emission, with an increment of 20 nm. The measurement time of each complete scan took 90 seconds. The spectra of each batch were later preprocessed by subtracting the first spectrum (after inoculation) and using a median filter. Also, the scattered light was excluded from the spectra. The three-dimensional data was a collection of EEM spectra of size 15 x 15 (15 excitation and 15 emission wavelengths) indexed by the time. From the total 225 possible locations on the 15 x 15 spectra, only 120 were valid pairs (no excitation can lead to emission with a higher energy – smaller wavelength). Fermentations P1, P2, and P3 had 412, 472, and 450 spectra, respectively, in a total of 1334 EEMs. As the data is based on subtracted spectra of the first spectrum, the measurement unit of the spectra is the relative fluorescence intensity. Figure 5.3 presents the absolute change of each Ex/Em pair's relative fluorescence intensity for the fermentations.

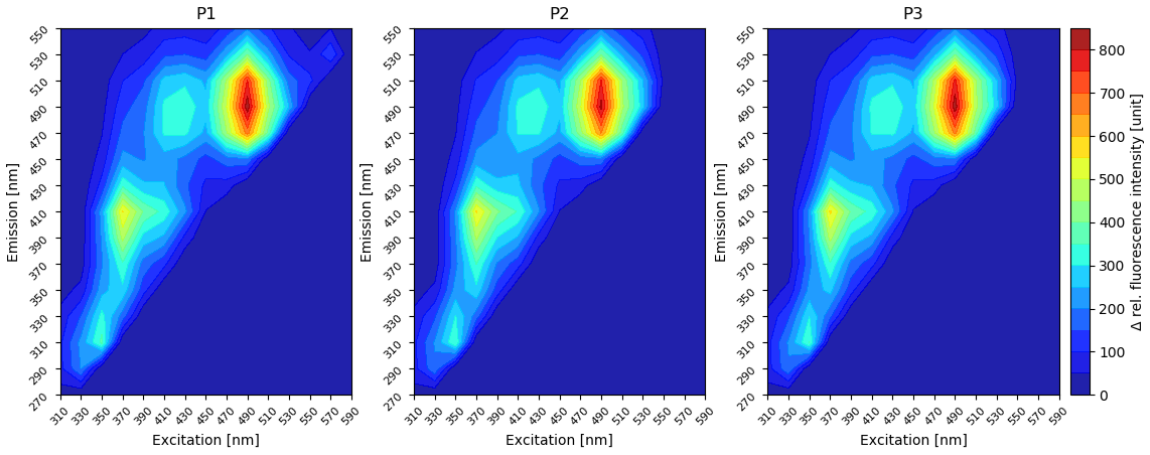



Figure 5.3. Absolute change in the relative fluorescence intensity of each Ex/Em pair, for each fermentation. P1, P2 and P3 represent each of the three batch fermentations.

For the offline analysis, samples were regularly taken from the bioreactor and centrifuged. The wet cells were left in a drying oven for 24 hours and weighted when cooled. The supernatant of the samples after centrifugation was examined by HPLC to determine glucose and ethanol concentration. The offline measurements were extrapolated to match the same 90 seconds' window. Figure 5.4 presents the offline and extrapolated measurements. All three batch runs were operated at the same conditions: 30 °C and pH 5. The detailed information about yeast strain, culture, fermentation conditions, and all the equipment used in the offline analysis can be seen in the referred work.

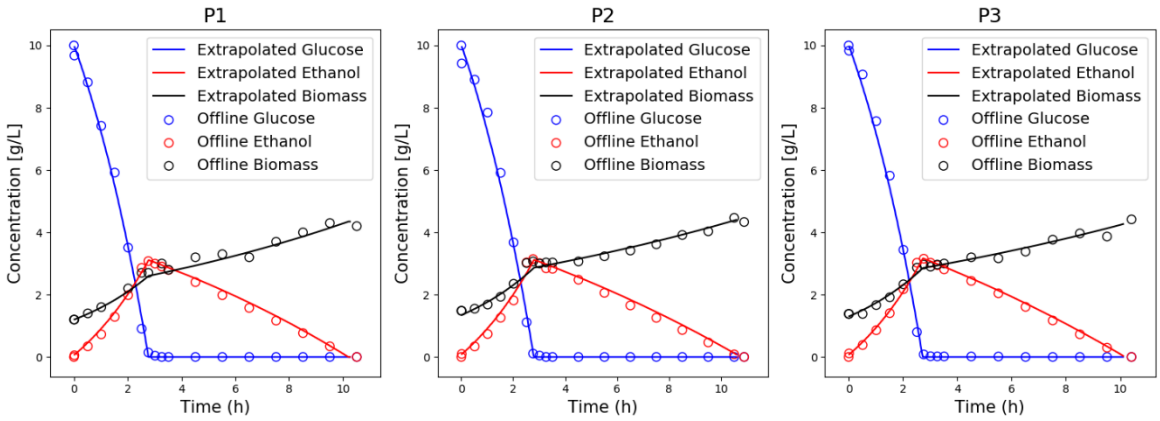



Figure 5.4. Offline and extrapolated measurements for the outputs of interest in each fermentation.

The data from the three fermentations were concatenated, each sample identified by its timestamp, containing the EEM spectra and the extrapolated biomass, ethanol, and glucose concentration. The dataset was divided into training, validation, and testing subsets based on the modified systematic sampling methodology proposed by Santos *et al.* (2019): the dataset was sorted by ascending order of biomass concentration. A repeatable pattern was adapted to split the dataset 60%-20%-20% training, validation, and testing subsets, respectively (train-train-train-val-test…), making sure both extreme values were contained in the training subset to avoid extrapolation.

### *5.2.2 Neural Networks*

All implementations in this work were performed in Python v.3.5. with Keras (high-level neural networks API capable of running on top of TensorFlow) to fit the neural networks.

#### *5.2.2.1 Deep Residual Network for Regression – R-ResNet*

The implemented residual network of this study is based on the code available in Keras documentation (keras.io/examples/cifar10_resnet/). The following section will explain the network architecture and its components.

**Convolutional layer – Conv2D.** It consists of a set of learnable filters of fixed dimensions. Each filter slides along the width and height of the input volume and computes dot products between its weights and the activation map of previous layers. How much the filter slides in each direction is defined by the stride (*e.g.,* a stride of 2 will cut in half the feature map).

**Average Pooling layer.** Down-sample the convolutional features using an average operation, summarizing features in patches of the feature map.

**Batch Normalization (BN) layer.** Introduced by Ioffe and Szegedy (2015), BN layers improve the network's speed, performance, and stability by mitigating the covariate shift when input distributions to a learning system changes. BN layers normalize the data in each mini-batch across the network and are often used after convolutional layers.

**Flatten layer.** Here, the convolutional map of shape (N x M x F) is converted to a simple vector of (N*M*F) shape.

**Fully-Connected (FC) layer.** Standard neural network layer where each neuron has full connections to all neurons in the previous layer.

**Overall Architecture.** The residual blocks and structure of the network can be seen in Figure 5.5. The residual blocks (RB) follow the work of He *et al.* (2016b): stacks of (1x1) – (3x3) – (1x1) batch normalization – ReLU – Conv2D layers (also known as residual bottleneck block). Stage 0 starts with a Conv2D – BN – ReLU without skip connections, where we set the initial number of filters. In Stage 1 we have the first RB with Projection Shortcut (RB – $PS_0$), where the feature map is kept the same and the filters are quadrupled. It is followed by $N_1$ stacks of RB with Identity Shortcut (RB – IS). Within each stage, the layers have the same number of filters and the same feature map sizes. Each Stage 2 and 3 begins with a RB – PS, where the feature map size is halved by a convolutional layer with stride 2, while the number of filter maps are doubled. They are followed by $N_2$ and $N_3$ stacks of RB – IS, respectively. Finally, in Stage 4, after BN – ReLU, an Average Pooling layer is used and flattened to a vector. The vector is connected to an FC layer with linear activation and to the desired output.

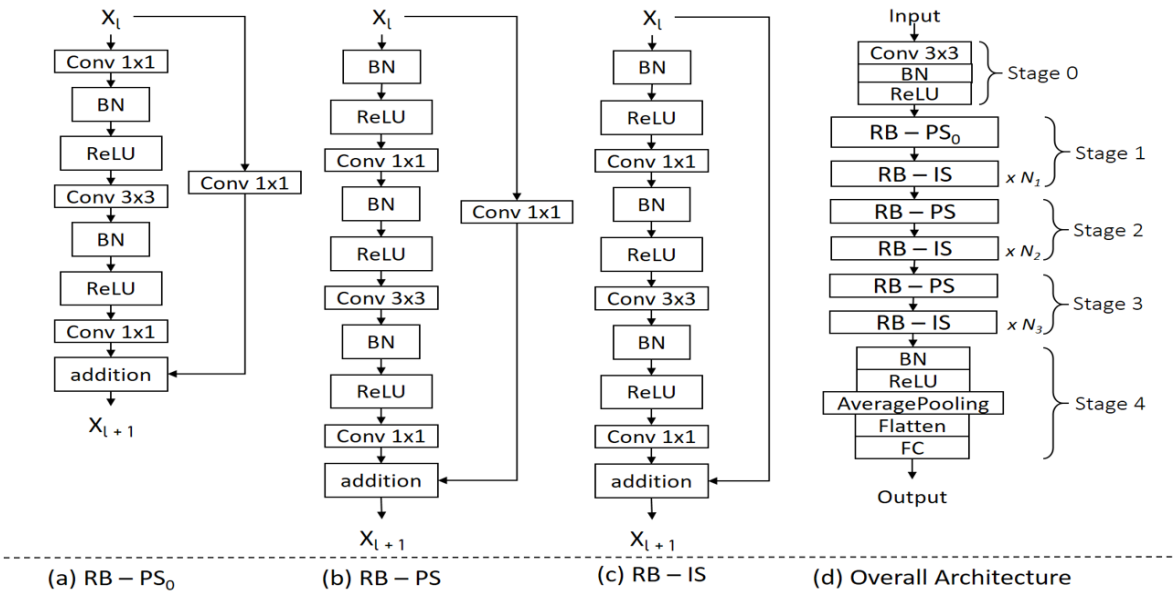



Figure 5.5. Residual blocks and overall architecture of the network. (a) First Residual Block with Projection Shortcut (RB – $PS_0$). (b) Residual Block with Projection Shortcut (RB – PS). (c) Residual Block with Identity Shortcut (RB – IS). (d) Overall Architecture of the network.

#### *5.2.2.2 Autoencoder*

The building of the Keras' AE can be guided by the schema presented in Figure 5.6, which highlights the main aspects to consider in the architecture, the loss function, and the activation functions.

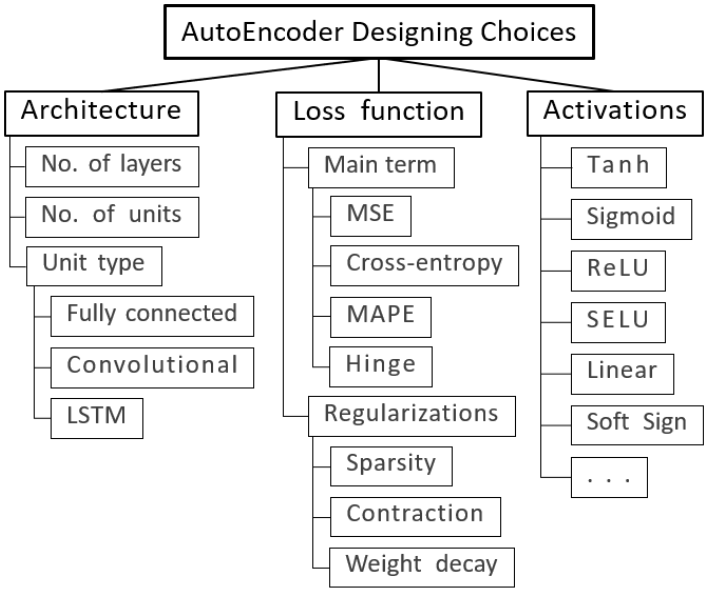



Figure 5.6. Summary of possible parameters of an autoencoder.

For the fitting of the AEs, each EEM spectra were transformed in a single row vector, the row being the time tag (representing one sample), and the columns being the fluorescence intensity of each valid Ex/Em pair (a total of 120 pairs). The input and output dimensions of the AE were, so, *N* x 120, where *N* is the number of samples.

The parameters of the autoencoder were decided based on a grid-search. The most straightforward architecture on the number of layers was applied: three layers, input – bottleneck – output. In the bottleneck, the use of 3 to 10 neurons was tested. Deng et al. (2014) mentioned in a similar case that the training of a neural network is a problem that has a computational complexity of the order of the product of the number of units in the hidden layers and the number of layers. Therefore, it is recommended only to escalate the architecture if more accuracy and effectiveness are needed.

The chosen unit type is the fully connected or Keras' dense, where every neuron connects with all neurons from the precedent and subsequent layers.

The loss function (the objective function to be minimized) can be chosen based on some knowledge of the data, and we selected the mean squared error – MSE – for this study.

For the activations of each layer, we study the use of all the non-classificatory ones available in Keras – ELU, SELU, ReLU, Tanh, Sigmoid, Softsign, Softplus, and Linear – to later compare the results of different combinations.

Neural network optimization methods are usually based on the stochastic gradient descendent (SGD) and its variants, like AdaDelta, RMSProp, and Adam. RMSProp with a learning rate of 0.0001 and decay of $1.10^{-6}$ was chosen for the minimization of the loss function.

Each AE was fitted using the training and validation subsets for 1000 epochs and a batch size of 64. The reconstruction RMSE for the training, validation, and testing subsets were evaluated and stored. If the quality metrics are similar within subgroups, one can expect that new samples, never seen before by the AE, should follow a similar reconstruction quality.

The selected AE was the one with the smallest pondered reconstruction RMSE for the training/validation/testing subsets, calculated as:

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(y_i - \hat{y}_i)^2}{N}} \tag{5.1}$$

where $y$ stands for the measured output value and $\hat{y}$ for the predicted output value.

### *5.2.3 Trust Screening and Anomaly Detection*

The fitted AE is applied as a tool for anomaly detection and trust screening: the reconstruction error of new samples using the trained AE is calculated. If the error is superior to a defined threshold, the sample cannot satisfy the standard correlation learned during the training phase, and it is considered abnormal. Here, the threshold is set as the greater reconstruction RMSE in the training/validation/testing subsets. Any new sample flagged as an anomaly should be treated with particular attention to avoid misleading predictions in soft sensors trained with "normal" data (here, the R-ResNet model). The higher the reconstruction error, the less the sample should be trusted.

A defect test was created to simulate measurement problems in input variables to evaluate the AE capability to abnormal flag samples. Some samples in each fermentation were selected, and error was added to one of their input features. The error consisted of a random number between -1 and 1 multiplied by the original value of that input in that sample. This way, a small number of defect samples were created, which could be very subtle (if the multiplier is close to zero, the input is almost unchanged) or more aggressively changed (if the multiplier is -1 that input will have zero intensity, and if it is 1, the input will be doubled). The defect dataset is evaluated with the trained AE, and the reconstruction RMSE is calculated. If the defect samples have errors higher than the threshold, they are considered abnormal. The trained R-ResNet model predictions are also evaluated using the original and the defect datasets to compare results.

The input feature to whom error is added was selected as pair Ex450/Em530, related to flavins. As can be seen in Assawajaruwan *et al.* (2017) research, the information contained in the EEM spectra related to biomass, ethanol, and glucose is mainly focused in regions that correspond to biogenic fluorophores (*e.g.,* NAD(P)H, tryptophan, pyridoxine, and flavins). Although the reconstruction error should detect a change in any relevant pair, if that pair is irrelevant for the R-ResNet model, it would not cause any disturbance in the predictions.

## 5.3 Results and Discussions

### *5.3.1 R-ResNet*

Following the structure presented in Figure 5.5 (d), the residual network parameters $N_1, N_2, N_3$, were all chosen as 8, being the network 83 layers deep. The average polling layer used a 4 x 4 kernel, and the FC layers had 3 neurons and linear activation, the same number of outputs of our system (biomass, ethanol, and glucose concentrations). Table 5.1 describes a summary of the structure.

Table 5.1. R-ResNet structure summary.

| Stage | Layer | Output size | Type of Convolution |
|---|---|---|---|
| - | Input | 15 x 15 | |
| Stage 0 | Conv2D | 15 x 15 x 16 | $[3\ x\ 3, 16]$ |
| Stage 1 | RB - $PS_0$ | 15 x 15 x 64 | $\begin{bmatrix} 1\ x\ 1, 16 \\ 3\ x\ 3, 16 \\ 1\ x\ 1, 64 \end{bmatrix}$ |
| | RB - IS | 15 x 15 x 64 | $\begin{bmatrix} 1\ x\ 1, 16 \\ 3\ x\ 3, 16 \\ 1\ x\ 1, 64 \end{bmatrix} x\ 8$ |
| Stage 2 | RB - PS | 8 x 8 x 128 | $\begin{bmatrix} 1\ x\ 1, 64 \\ 3\ x\ 3, 64 \\ 1\ x\ 1, 128 \end{bmatrix}$ |
| | RB - IS | 8 x 8 x 128 | $\begin{bmatrix} 1\ x\ 1, 64 \\ 3\ x\ 3, 64 \\ 1\ x\ 1, 128 \end{bmatrix} x\ 8$ |
| Stage 3 | RB - PS | 4 x 4 x 256 | $\begin{bmatrix} 1\ x\ 1, 128 \\ 3\ x\ 3,128 \\ 1\ x\ 1, 256 \end{bmatrix}$ |
| | RB - IS | 4 x 4 x 256 | $\begin{bmatrix} 1\ x\ 1, 128 \\ 3\ x\ 3,128 \\ 1\ x\ 1, 256 \end{bmatrix} x\ 8$ |
| Stage 4 | Average Pooling | 1 x 1 x 256 | $[4\ x\ 4]$ |
| | Flatten | 256 x 1 | |
| | FC | 3 x 1 | |
| | Output | 3 | |

To avoid initialization problems, the R-ResNet was fitted multiple times, discarding any run that did not converge. The networks were fitted with the training and validating subsets for 150 epochs, using Adam optimizer with a learning rate of 0.001 and the validation mean squared error as the loss function. After fitted, the regression model was used to predict biomass, ethanol, and glucose concentration for all subsets, and the RMSE and the coefficient of determination ($R^2$) were calculated. Table 5.2 compiles the metrics of the best fitted R-ResNet.

Table 5.2. R-ResNet prediction metrics, relative to the extrapolated offline measurements.

| | $R^2$ | | | RMSE [g/L] | | |
|---|---|---|---|---|---|---|
| | Training | Validation | Testing | Training | Validation | Testing |
| Biomass | 0.9993 | 0.9992 | 0.9992 | 0.02196 | 0.02344 | 0.02325 |
| Ethanol | 0.9994 | 0.9992 | 0.9993 | 0.02186 | 0.02463 | 0.02410 |
| Glucose | 0.9999 | 0.9998 | 0.9998 | 0.03189 | 0.03865 | 0.03717 |

As shown in Table 5.2, the regression model could predict the outputs very satisfactorily, with almost perfect accuracy. Figure 5.7 shows the predicted versus measured outputs during the fermentations. The combination of EEM fluorescence spectroscopy and regression with residual networks proved to be a powerful tool for tracking fermentation bio-products.

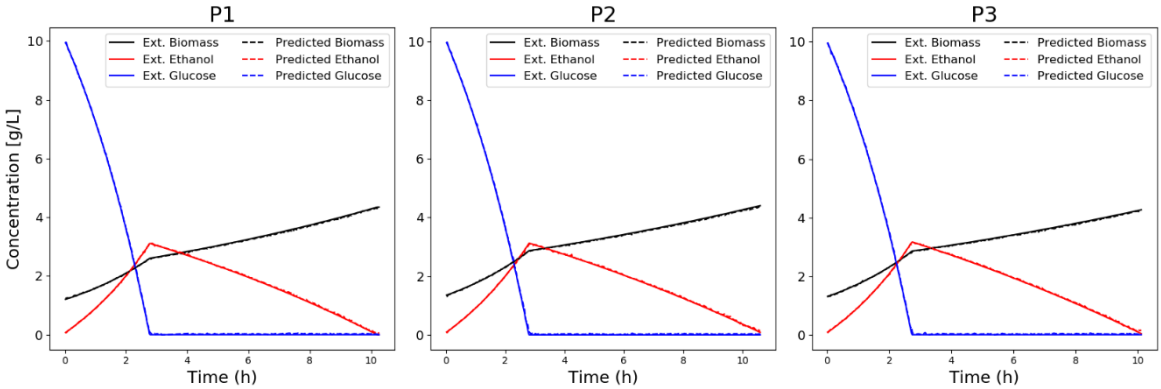



Figure 5.7. R-ResNet predicted and extrapolated measured outputs for P1, P2, and P3 fermentations. For visualization purposes, data points were linearly connected as continuous lines.

### *5.3.2 Autoencoder*

The best-fitted AE had ReLU activation in the encoder layer, linear activation in the decoder, and 3 neurons in the bottleneck layer (this means a reduction of dimensionality from 120 – original input – to 3). More neurons in the bottleneck did not make a significant improvement in the AE error. The metric chosen to compare the AEs, in this case, was the RMSE. Figure 5.8 presents the autoencoder reconstruction RMSE ordered by the time during the fermentations.

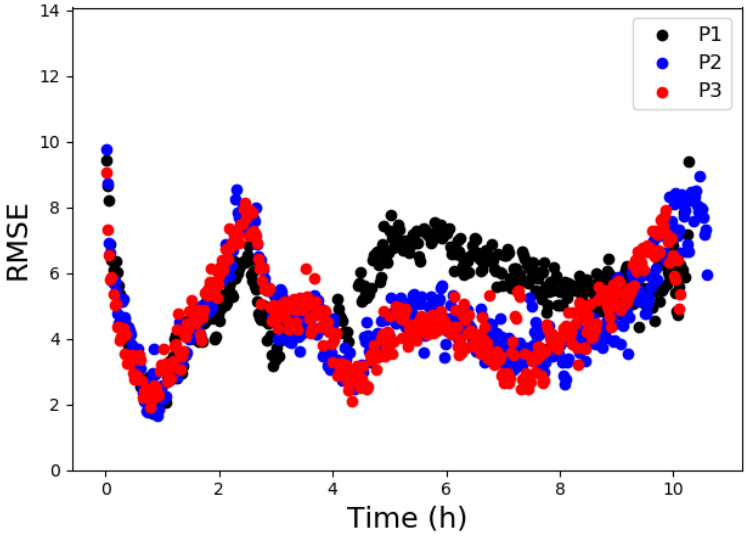



Figure 5.8. Autoencoder predictive reconstruction root mean squared errors for P1, P2, and P3 fermentations.

The mean reconstruction RMSE of each fermentation was 5.46, 4.70, and 4.57 rel. fluorescence intensity units, related to P1, P2, and P3 (a general error lower than 0.65%, considering that fermentation P1, P2, and P3 had an absolute maximum change of relative fluorescence intensity of, respectively, 858, 861, and 858 units). The greatest RMSE among fermentations was close to 10 rel. Fluorescence intensity units. Then, the threshold was defined: if the RMSE of any new sample was greater than 10, it would be flagged as abnormal and should be treated with care. The higher the reconstruction error, the less that new sample resembles the training data and less trust. The fitted R-ResNet regression model could predict the output concentrations within the same range of expected errors presented in Table 5.2.

3.3 Trust Screening and Anomaly Detection

To evaluate the AE capability to flag ill-suited new samples, we ran the defect test. For each fermentation, 5% of the original samples were selected to receive errors in the Ex450/Em530 fluorescence pair. The trained AE is used to reconstruct the new defect dataset. The reconstructions RMSE are evaluated, and the biomass, ethanol, and glucose concentrations are predicted using the fitted R-ResNet model. Figure 5.9 compiles the results for the trust screening study. We can see the R-ResNet predicted outputs for all three fermentations in the superior portion of the graphics. The inferior graphics present the calculated AE reconstruction error for the defect dataset. Samples marked with a dot that received errors and the dot location represent the output prediction. Also, the percentage of the original intensity added to each sample is shown.

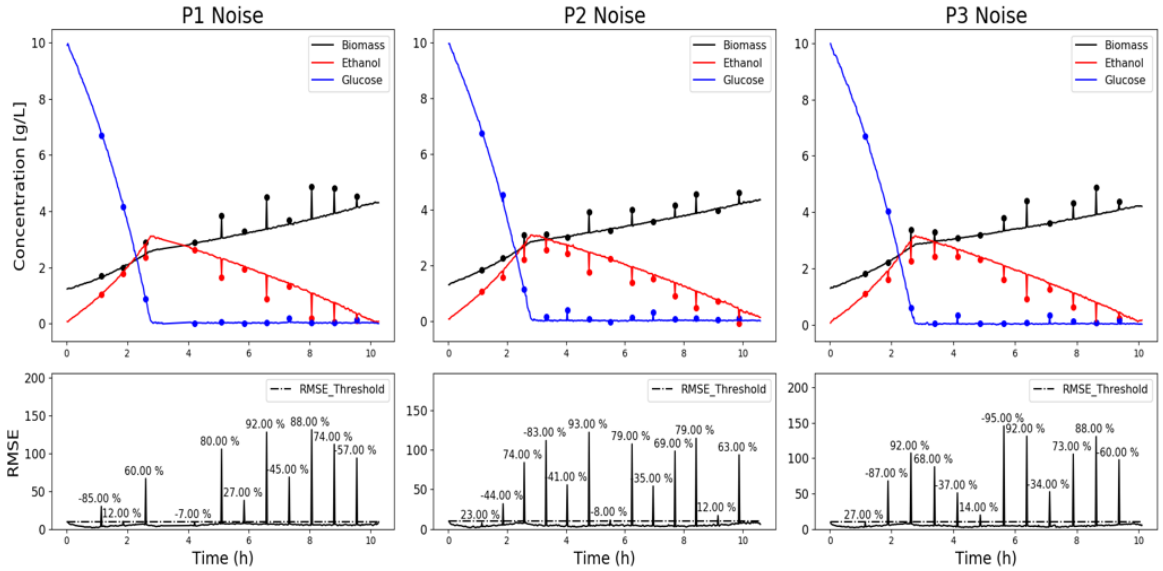



Figure 5.9. Defect test results. The superior graphics present the R-ResNet output predictions for each fermentation, and the inferior graphics the AE reconstruction RMSE. Samples marked with a dot are the ones that received fluorescence intensity errors in pair Ex450/Em530. The percentage of the original intensity added to each sample is described. For visualization purposes, data points were linearly connected as continuous lines.

Analyzing the results, we can see that the AE RMSE could flag all samples that received significant errors. Some samples got none or very little intensity added, and this can be seen in the market samples where the RMSE was almost unchanged (as expected).

Figure 5.10 presents the reconstruction errors for all samples, with a constant 15% added intensity. In this case, more than 85% of samples would be flagged as abnormal. As expected, the non-flagged samples are at the beginning of the fermentations, where fluorescence intensities are smaller, and the squared errors are less significant overall.

For the biomass, ethanol, and glucose concentrations, almost every defect sample presented higher prediction errors. Nevertheless, not all samples with RMSE higher than the threshold presented a negative change in the output predictions (*e.g.,* around 6 hours in P1). The R-ResNet model was robust to a fair amount of added errors without loss of prediction quality. This robustness can be explained by the nature of the EEM fluorescence data and how CNN treats the spatial input: the information in the spectra is not punctual but dispersed in neighboring pairs, and the convolutional filters learn this multi-input information of adjacent inputs.

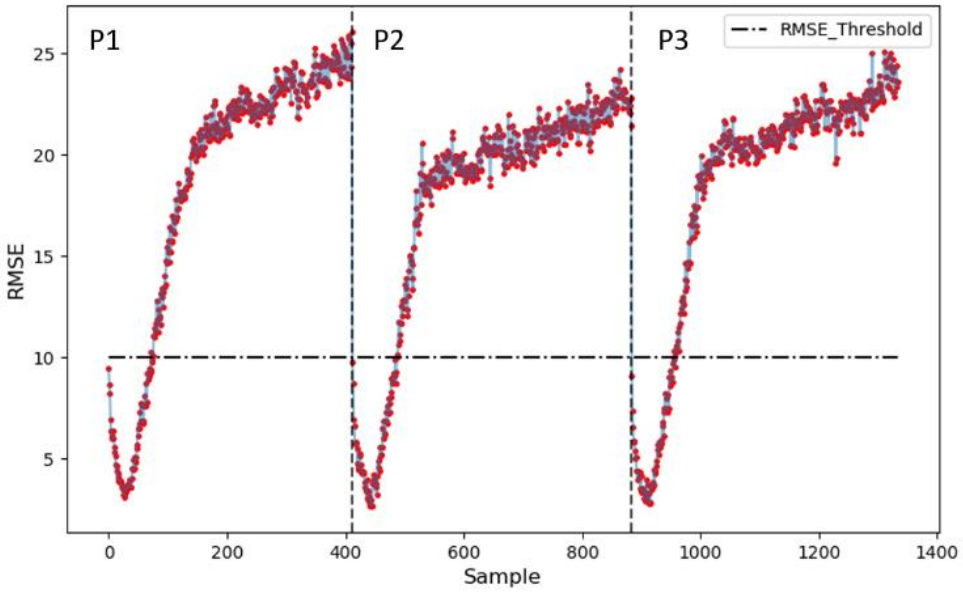



Figure 5.10. Autoencoder reconstruction RMSE for all samples with a constant 15% added relative fluorescence intensity.

As proposed in our methodology, the higher the reconstruction error, the less the new sample should be trusted, and the greater the chance of misleading predictions. Taking as an example the fermentations presented here, if the system would be controlled based on *online* fluorescence measurements, it would be wiser not to take any controlling actions based on readings flagged as abnormal by the AE, and investigate the cause. The AE threshold can be tuned to reflect the amount of error (trust) accepted by the operator. If a sequence of new samples presents high reconstruction errors, that can be a strong indication of a problem in collecting data (equipment fault) or that the system is no longer well represented by the training data.

## 5.4 Conclusions

Fluorescence spectroscopy is a valuable tool for monitoring cultivation processes, especially for autofluorescence microorganisms like yeast. The combination of EEM data and convolutional neural networks has become the state-of-art for many predictive soft sensors. Although the use of black-box models is widely spread in the industry, it is only possible to ensure the quality of their predictions in the limited range where it was trained. There is no definitive methodology to evaluate if a fitted black-box model could satisfactorily predict a new data point of which no characteristics are known. In this work, we proposed the use of deep residual convolutional neural networks and EEM fluorescence to predict bioprocess variables in *S. cerevisiae* fermentation and the use of autoencoder reconstruction errors as a trust screening of new samples.

For the *S. cerevisiae* fermentation, the regression ResNet model predicted biomass, ethanol, and glucose concentrations very satisfactorily, with $R^2$ greater than 0.99 and root mean squared errors of prediction of 0.02325, 0.02410 and 0.03717, respectively, for the testing subgroup.

The best-fitted autoencoder had a 3-layer architecture, with three neurons in the bottleneck and using ReLU activation for the encoder and linear activation for the decoder.

The mean reconstruction RMSE for the three fermentations was 4.91 rel. Fluorescence intensity units, representing a general error smaller than 1%.

To evaluate the AE capability to flag ill-suited new data points, random fluorescence intensity was added (or subtracted) from the Ex450/Em530 pair (related to flavins) in some samples, creating a defective dataset. The dataset was evaluated with the trained AE and the R-ResNet model to compare reconstruction errors and bioprocess concentrations. The AE was able to flag samples with significant added errors, and, as expected, the defective samples also presented higher predictive errors in general. Given the nature of the EEM fluorescence data and how the convolutional neural network learns information, the R-ResNet model was robust to a fair amount of added errors without losing predictive quality. Even though not all flagged samples presented a relevant increase in prediction errors, the higher the reconstruction RMSE, the less the new sample should be trusted to avoid misleading predictions and erroneous controlling actions.

The proposed methodology was considered successful both as a tool to predict bioprocess variables, as a trusted screening to avoid misleading predictions in soft sensors/ black-box models.

## 5.5 References

ALMOTIRI, J.; ELLEITHY, K.; ELLEITHY, A. Comparison of Autoencoder and Principal Component Analysis Followed by Neural Network for E-Learning Using Handwritten Recognition. **2017 IEEE Long Island Systems, Applications and Technology Conference, LISAT 2017**, n. May, 2017.

AMARBAYASGALAN, T.; JARGALSAIKHAN, B.; RYU, K. Unsupervised Novelty Detection Using Deep Autoencoders with Density Based Clustering. **Applied Sciences**, v. 8, 2018.

ASSAWAJARUWAN, S.; REINALTER, J.; HITZMANN, B. Comparison of methods for wavelength combination selection from multi-wavelength fluorescence spectra for on-line monitoring of yeast cultivations. **Analytical and Bioanalytical Chemistry**, v. 409, n. 3, p. 707–717, 2017.

CHARTE, D. et al. A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. **Information Fusion**, v. 44, n. December 2017, p. 78–96, 2018.

DENG, J. et al. Autoencoder-based unsupervised domain adaptation for speech emotion recognition. **IEEE Signal Processing Letters**, v. 21, n. 9, p. 1068–1072, 2014.

DOERSCH, C. Tutorial on Variational Autoencoders. p. 1–23, 2016.

FAASSEN, S. M.; HITZMANN, B. Fluorescence spectroscopy and chemometric modeling for bioprocess monitoring. **Sensors (Switzerland)**, v. 15, n. 5, p. 10271–10291, 2015.

HANIF, M. S.; BILAL, M. Competitive residual neural network for image classification. **ICT Express**, v. 6, n. 1, p. 28–37, 2020.

HE, K. et al. **Deep Residual Learning for Image Recognition**. [s.l: s.n.].

HE, K. et al. **Identity Mappings in Deep Residual Networks**. [s.l: s.n.]. v. 9908

HE, K.; SUN, J. Convolutional Neural Networks at Constrained Time Cost. 2014.

HU, Y. et al. **Spiking Deep Residual Network**. [s.l: s.n.].

IOFFE, S.; SZEGEDY, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015.

ITAKURA, K. et al. Estimation of Citrus Maturity with Florescence Spectroscopy Using Deep Learning. **Horticulturae**, v. 5, n. 1, p. 2, 26 dez. 2018.

JES, F. et al. Deep Convolutional Autoencoders vs PCA in a Highly-Unbalanced Parkinson ' s Disease Dataset : A DaTSCAN Study. **Springer Nature**, p. 47–56, 2019.

LATHUILIÈRE, S. et al. A Comprehensive Analysis of Deep Regression. 2018.

MANNING-DAHAN, T. PCA and Autoencoders. 2017.

MASIERO, S. S. et al. Evaluation of wavelength selection methods for 2D fluorescence spectra applied to bioprocesses characterization. **Brazilian Journal of Chemical Engineering**, v. 30, p. 289–298, 2013.

MASLANKA, R.; KWOLEK-MIREK, M.; ZADRAG-TECZA, R. Autofluorescence of yeast Saccharomyces cerevisiae cells caused by glucose metabolism products and its methodological implications. **Journal of Microbiological Methods**, v. 146, p. 55–60, 2018.

PAQUET-DURAND, O. et al. Artificial neural network for bioprocess monitoring based on fluorescence measurements: Training without offline measurements. **Engineering in Life Sciences**, v. 17, n. 8, p. 874–880, 9 out. 2017.

PODRAZKÝ, O. et al. Monitoring the growth and stress responses of yeast cells by two-dimensional fluorescence spectroscopy: First results. **Folia Microbiologica**, v. 48, n. 2, p. 189–192, 2003.

RHEE, J. IL; KANG, T.-H. On-line process monitoring and chemometric modeling with 2D fluorescence spectra obtained in recombinant E. coli fermentations. **Process Biochemistry**, v. 42, n. 7, p. 1124–1134, 2007.

RUTHERFORD, J. W. et al. Excitation emission matrix fluorescence spectroscopy for combustion generated particulate matter source identification. **Atmospheric Environment**, v. 220, p. 117065, 2020.

SAKURADA, M.; YAIRI, T. Anomaly detection using autoencoders with nonlinear dimensionality reduction. **Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis**, p. 4, 2014.

SANTOS, P. V. J. L. et al. K-RANK: AN EVOLUTION OF Y-RANK FOR MULTIPLE SOLUTIONS PROBLEM. **Brazilian Journal of Chemical Engineering**, v. 36, p. 409–419, 2019.

SILVEIRA, D.; COELHO, T.; SANTOS, A. Evolution of Black-Box Models Based on Volterra Series. **Journal of Applied Mathematics**, v. 2015, 2015.

XUE, Y. et al. **Cell Counting by Regression Using Convolutional Neural Network**. (G. Hua, H. Jégou, Eds.)Computer Vision – ECCV 2016 Workshops. **Anais**...Cham: Springer International Publishing, 2016

YAMSHCHIKOV, I. P.; TIKHONOV, A. Music generation with variational recurrent autoencoder supported by history. 15 maio 2017.

ZEILER, M. D.; FERGUS, R. **Visualizing and Understanding Convolutional Networks**. (D. Fleet et al., Eds.)Computer Vision – ECCV 2014. **Anais**...Cham: Springer International Publishing, 2014

# Capítulo 6 – **Developing Impurity Soft Sensors for a Propylene/Propane Splitter Unit Using Stochastic Variable Selection**

**Abstract:** Propylene is a key intermediate in the petrochemical industry. Historically a by-product of cracking, the gap in supply and demand motivated the on-purpose production, as the recovery from liquefied petroleum gas. Irrespective of the manufacturing route, the product is always a mixture of propylene/propane that must be separated to meet purity grades. This article proposes the use of stochastic metaheuristics and regularization techniques for the development and optimization of empirical predictive models, capable of inferring impurities and other essential process variables, from the polynomial expansion of readily available information in a Propane/Propylene Separation Unit. The studied Unit was composed of three separation columns. For Column T-01, the main objective is predicting the mass ratio of the heavy hydrocarbons ($Z_{C4+}^{D1}$) impurity in the distillate stream. For samples with $Z_{C4+}^{D1}$ smaller than 0.01 (more relevant to the operation of the Unit), the AnTSbe model kept 11 out of the 219 available inputs, with RMSE of 2.1e-5, MAPE of 2.14%, and Maximum Absolute Percentage Error (Max e%) of 7.40%. It was also possible to predict the concentrations of ethane – $Z_{C2}^{D1}$, propylene – $Z_{C3-}^{D1}$ and propane – $Z_{C3+}^{D1}$ into the distillate stream satisfactorily, with Max e% smaller than 3%. For Column T-02, it is economically essential to reduce the waste of propylene in the distillate stream ( $Z_{C3-}^{D2}$). The AnTSbe model kept 32 out of the 454 variables and was able to predict $Z_{C3-}^{D2}$ with RMSE of 0.0162, MAPE of 2.11%, but a high Max e% of 45.6%. Evaluating the distribution of the errors, models were less precise in higher concentrations of propylene, not relevant for Unit operation. In Column T-03 it is crucial to maintain the specification of 0.004 kg/kg impurity of propane in the distillate stream( $Z_{C3+}^{D3}$). For concentrations of propane lower than 0.01 kg/kg, the AnTSbe model kept 17 out of the 1539 available variables and predicted the impurity of interest with RMSE of 2.1e-5, MAPE of 83%, and Max e% of 2910%. A meticulous evaluation of the individual errors showed that for the range of mass ratio between 0.0002 and 0.01 (region of most significant interest for the operation of the Unit), the MAPE error is lower than 5% and the Max e% is 16%.

## 6.1 Introduction

Light olefins are essential to the petrochemical industry because they are the basic building blocks for many end products, as polyethylene and polypropylene. In this regard, propylene is considered one of the key intermediate in the industry, with derivatives that include acrylic acid, acrylonitrile, cumene, phenol, alkylate, high-octane gasoline blends, trimmers, and tetramers for detergents (HEINRITZ-ADRIAN; WENZEL; YOUSSEF, 2008). Propylene is commonly produced as a byproduct of steam cracking (SC) and fluidized catalytic cracking (FCC) of naphtha and light gas oil, with gasoline and ethylene being the main products. Propylene production depends upon the feedstock and the operational rates. When heavy liquid cracking is performed, propylene is readily available. However, most modern steam crackers use ethane-based feed, leading to less propylene being produced, causing a gap between supply and demand (BROOKS, 2013). To meet propylene market demand, refineries are gradually adopting technologies for propylene production, propane dehydrogenation, olefin metathesis, and methanol to propylene (DIMIAN; BILDEA; KISS, 2019; HEINRITZ-ADRIAN; WENZEL; YOUSSEF, 2008). Irrespective of the propylene manufacturing route, the product is always a mixture of propylene/propane that must be separated to meet purity grades: for polymer grade, a minimum of 99.5% of purity is required, the chemical grade is 90 – 95%, and refinery grade 50 – 70% (BRYAN, 2004).

Liquefied Petroleum Gas (LPG), produced in natural gas processing facilities from underground formations or in refinery operations (mostly FCC), is a mixture of propane, butylenes, and many other hydrocarbons often containing relevant amounts of propylene. LPG is mainly used as a fuel, and propylene, when burned, can cause problems, leaving deposits in engines and injectors. It can also polymerize in storage tanks or fuel lines, which is a particular concern when LPG is used as a standby fuel source. The current American specification, HD-5, on fuel-grade propane limits the propylene content to 5% (GPSA, 2012). Economically, polymer grade propylene is highly sought-after, reaching selling prices 2 to 3 times higher than LPG (ECHEMI, 2020). There were continuous efforts in the past decade to improve and maximize propylene production from FCC processes (AKAH; AL-GHRAMI, 2015) to recovery propylene from LPG (VENKATESH BABU; RAMESH, 2013) and to enhance propylene production by other routes (LAVRENOV et al., 2015). Palmer *et al.* (2012) presented a detailed study and financial analysis for investment opportunities for refiners predicting the future high price of propylene, concluding that making high purity propylene from LPG using propane/propylene splitter units was economically viable and encouraged.

The separation of propane and propylene is challenging, considering they have similar molecular sizes and physical properties. Oppositely to the usual distillation operations, the difference in temperatures inside the distillation column does not aggregate as much information and cannot be easily linked to product purity. The process requires high capital cost and high energy consumption, as the reflux ratio needs to be kept high for maximum separation (UMO; BASSEY, 2017).

To keep the process profitable, simulations of separation units are often explored, trying to establish optimum operating parameters for better operation with higher product quality and improved energy consumption (MAUHAR; BARJAKTAROVIC; SOVILJ, 2004; UMO; BASSEY, 2017). Simulations can also be the base for the development of control strategies, as the use of feedback and feedforward control (PĂTRĂŞCIOIU; ANH; POPESCU, 2015), Model Predictive Control (HINOJOSA; CAPRON; ODLOAK, 2016), and self-optimizing

control (SCHULTZ, 2015). As a more simplified approach, historical (or simulated) data can be used to develop data-driven black-box models that transform easily to measure information into challenging to obtain features (as concentration profiles), that are usually quantified by sampling from the Unit streams and have considerable delay (GRAZIANI; PAGANO; XIBILIA, 2010). The information provided by a soft-sensor can guide process engineers during daily operations, can be used as a base for controlling actions, check the instrumentation system, and be constantly updated using new offline measurements.

In this work, we propose a methodology to develop black-box impurity soft-sensors for a propylene/propane separation unit, based on the polynomial expansion of readily available features from the process, applying stochastic metaheuristics and regularization techniques for variable selection and optimization of the predictive models.

## 6.2 Materials and Methods

### *6.2.1 Propylene/propane Splitter Unit (PPSU)*

The dataset used in this study was developed by Schultz (SCHULTZ, 2015), where an actual propylene/propane splitter unit in operation was used as inspiration for Aspen Plus™ and neural network simulations of the process to study self-optimizing control (SOC) techniques.

The main goal of the unit is to produce high purity (99.6%) propylene stream (C3-) from an input stream of liquefied petroleum gas (LPG). The process consists of three main distillation columns in series, been the LPG feed to the first column (T-01). In T-01, heavy compounds (C4+) are removed from the bottom, and the propylene-rich distillate stream feeds the second column (T-02). From T-02, the distillate stream rich in ethane (C2) is extracted, and the bottom stream, containing mainly propane (C3+) and propylene (C3-), feeds the third column (T-03). At last, high purity propylene is withdrawn from T-03 distillate stream. A simplified flowchart of the splitter unit can be seen in Figure 6.1.

The stationary model of the unit was simulated in Aspen Plus™ version 7.2 with the use of Peng-Robinson's thermodynamic model to calculate the physical-chemical proprieties of the streams. The complete set of the simulation parameters can be seen in the referred work. The degrees of freedom of each column were defined as for T-01, the reflux ratio ($RR_1$) and the mass ratio between the distillate stream flow rate and the input flow rate (D/F1); For T-02, the reflux ratio ($RR_2$) and the mass ratio between the bottom stream flow rate and the column's input flow rate (B/F2); for T-03, the ratio of the flow that leaves the compressor and will be used as heating fluid for the reboiler (FA3), and the ratio of the flow that leaves the reboiler and return as reflux to the column (FR3). The simulation of the unit proved to have slow convergence, especially for T-03. As the volume of data needed for the study was significant, Schultz (SCHULTZ, 2015) used the Aspen results to train black-box models using neural networks (NN) to represent the system satisfactorily.

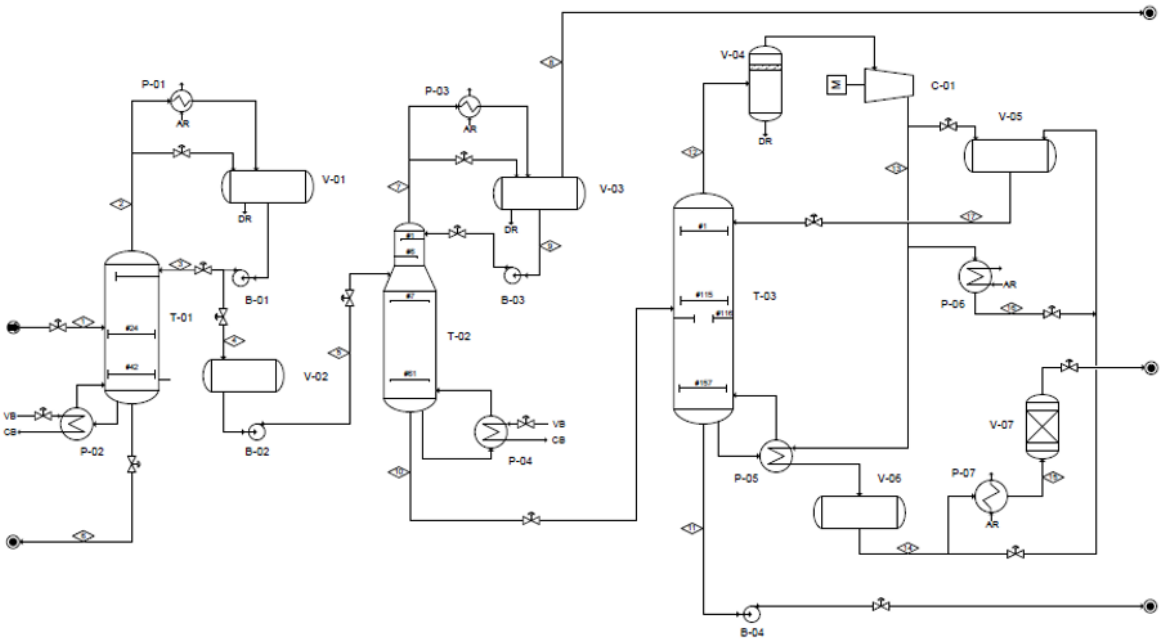



Figure 6.1. Simplified flowchart of the studied propane-propylene splitter unit. Source: (SCHULTZ, 2015).

The PPSU was divided into three separate systems, each represented by one of the distillation columns, and approximated by two neural networks: both were structured as dense two-layer fully-connected networks, using sigmoidal tangent as the activation function of the first layer, but one NN was focused on predicting the concentration profile of the output streams, also using sigmoidal tangent as the activation of the second layer, and the other NN was focused on predicting the remaining variables (*e.g.,* mass flow rates, temperatures, pressures), using a linear activation function. After training and testing the NNs satisfactorily with Aspen data, each column was simulated separately with a projected set of input data to obtain predictions covering the entire range of the operational region. Description of the inputs and their ranges for each column's NN follows.

For T-01, the dataset was constructed with the variation of the unit disturbs (feed mass flow rate $F_1$ and propane mass ratio in the feed stream $Z_{C3\text{-}}^{1}$) and the column's degrees of freedom ($RR_1$ and $D/F_1$). For T-02, six inputs were considered: feed mass flow rate ($F_2$), ethene mass ratio ($Z_{C2}^{2}$), propane mass ratio ($Z_{C3+}^{2}$), propylene mass ratio ($Z_{C3\text{-}}^{2}$), reflux ratio ($RR_2$), and mass ratio between input flow rate and bottom stream flow rate ($B/F_2$). The limits for T-02 input mass ratios and feed flow rates were obtained from the outputs of the T-01 simulations. The inputs for the T-03 model were feed mass flow rate ($F_3$), mass propylene ratio in the feed ($Z_{C3\text{-}}^{3}$), and the degrees as mentioned earlier of freedom $FA_3$ and $FR_3$. Also, for this case, the ranges for mass ratios and flow rates were derived from the outputs of T-02 simulations. The variables, nominal values, ranges, and the number of equidistant points for each simulation input are shown in Table 6.1.

With these arrangements, the full dataset consists in 900 operational points for T-01, 414720 for T-02 and 1764 for T-03.

Table 6.1. Description of the variables, nominal values, ranges, and the number of equidistant points used as inputs for each distillation column's neural networks' simulations. Source: (SCHULTZ, 2015).

| Column | Variable | Nominal Value | Range | N. of points |
|---|---|---|---|---|
| T-01 | $F_1$ | 63000 | 59850 - 66150 | 5 |
| | $Z_{C3-}^{1}$ | 0.355 | 0.319 - 0.390 | 6 |
| | $RR_1$ | 1.98 | 1.00 - 4.00 | 6 |
| | $D/F_1$ | 0.459 | 0.250 - 0.650 | 5 |
| T-02 | $F_2$ | 30064 | 11656 – 51630 | 10 |
| | $Z_{C2}^{2}$ | 0.0460 | 0.0171 – 0.109 | 8 |
| | $Z_{C3-}^{2}$ | 0.765 | 0.449 – 0.941 | 9 |
| | $Z_{C3+}^{2}$ | 0.188 | 0.0575 – 0.320 | 9 |
| | $RR_2$ | 12.23 | 8.00 – 15.00 | 8 |
| | $B/F_2$ | 0.896 | 0.600 – 0.910 | 8 |
| T-03 | $F_3$ | 26458 | 15874 – 37041 | 7 |
| | $Z_{C3-}^{3}$ | 0.798 | 0.594 – 0.953 | 7 |
| | $FA_3$ | 0.873 | 0.600 – 0.890 | 6 |
| | $FR_3$ | 0.792 | 0.642 – 0.942 | 6 |

### *6.2.2 Development of soft-sensors for the PPSU*

Although it was possible to simulate the PPSU satisfactorily using both Aspen™ and neural networks, this approach is very time-consuming, complicated, and usually not practical enough to be used routinely by process engineers as a tool for control or optimize the daily unit operation.

In this regard, the data generated by simulations can be used to develop soft-sensors, typically simple black-box linear (or non-linear) models that, once fitted, can readily transform available and easy to measure proprieties of the system in hard to obtain information, as concentration profiles (PAQUET-DURAND et al., 2017). They can also be used as a secondary tool to constantly audit the instrumentation of the unit once faulty sensors or wrong readings result in unexpected model predictions compared to offline/laboratory measurements (VENKATASUBRAMANIAN et al., 2003). Furthermore, the original input data can be expanded using mathematical operations (*e.g.,* logarithmical, exponential, or polynomial transformation of the inputs) to improve the model's capability to represent the process faithfully.

The methodology chosen to fit and optimize the predictive models is an adaptation of the AnTSbe algorithm proposed by Ranzan *et al.* (RANZAN; TRIERWEILER; TRIERWEILER, 2020): a compilation of tools to apply stochastic variable selection and linear model optimization to multidimensional data. The core of the algorithm is based on Ant Colony Optimization (STÜTZLE; LÓPEZ-IBÁÑEZ; DORIGO, 2011), where multiple parallel agents

(*ants*) select different combinations of the available inputs to minimize the objective function. The selection of inputs by the ants is guided by a random trigger associated with a quality indicator called *pheromone*. At the end of each iteration, the ants deposit pheromones in their selected variables inversely proportional to the error the model had in predicting the output of interest. The higher the pheromone an input has, the greater the chance it will be selected in future iterations. The algorithm is hybridized with Tabu Search (GLOVER, 1989), with the primary purpose of preventing those recent combinations of inputs from being reselected by ants in subsequential iterations, avoiding early stagnation. After the variable selection stage, each combination is used to fit linear models (with the option to apply regularization techniques, such as Ridge Regression – *l*2 norm, and Lasso Regression – *l*1 norm (TIBSHIRANI, 1994)), their quality metrics are calculated, and the model best fitted to minimize the objective function is updated. At the end of the optimization routine, the best-fitted model is presented, along with its selected variables and all calculated metrics. The selection of variables is crucial to improving model quality, especially in situations where there are highly correlated inputs or even inputs that do not correlate with the desired output (XU; ZHANG, 2001).

The methodology proposed in this work starts with the preprocessing of the data. All implementations in this work were done in Python v3.5.4.1 combined with the readily available modules, especially from the SciKit Learn library version 0.20.2 54 (PEDREGOSA et al., 2011).

First, the available data is divided into inputs and outputs to represent better the variables we want to predict, with only the inputs being part of the predictive models and each model with only one output (*MISO* – multiple inputs single output). In this work, all outputs will be mass ratios of the essential constituents of the PPSU streams (ethane – $C_2$, propylene – $C_{3-}$, propane – $C_{3+}$, and heavier hydrocarbons – $C_{4+}$).

Then, the inputs are normalized using the *StandardScaler* function from the Python SciKit Learn preprocessing library (PEDREGOSA et al., 2011), resulting in data with mean zero and unit standard deviation. This step is essential for using many Python linear models' libraries to be used throughout the algorithm.

Outliers on the dataset are detected using Hotelling's $T^2$ Statistic (HOTELLING, 1933) and removed. When working with empirical models, outliers can be very detrimental to model robustness and quality, masking essential patterns in the data. Before been used as inputs for the variable selection and model optimization routine, it is indispensable to have a clean, normalized and consistent dataset.

To further explore the potential use of the available information to predict the desired outputs, the inputs of each column (T-0x) will be expanded as the stacking of second ($x_i \cdot x_j$) and third-order ($x_i \cdot x_j \cdot x_k$) variables, with $i = j = k = 1, \dots, n$ (being *n* the number of original inputs). The new inputs will be regarded simply as new variables. Nevertheless, the fitted models will still be linear in the parameters.

The last preprocessing step is splitting the data into calibration (cal) and testing (test) subsets. The methodology selected for this is the one implemented by Santos *et al.* (SANTOS et al., 2019), being helpful to deal with multiple solutions problems (where multiple combinations of the input variables can yield the same output y). A k-means algorithm (RASCHKA, 2015) with *k* centroids (chose by the user) is run using only the input

variables to split the dataset into $k_i$ similar groups. For each cluster: the $k_{i,samples}$ are sorted in ascending order for a selected output y; the proportions of each subset are chosen (here, 66% calibration – 33% testing), and the methodology adapts a pattern to select, in order, the samples to their respective subsets (*e.g.,* cal-cal-test…). In this implementation, the extremes samples (with minimum and maximum values of y) in each cluster are always selected for the training subset to avoid extrapolation.

With the preprocessed data, the model fitting and optimization can begin. The calibration subset is passed as input for the AnTSbe algorithm, which will optimize the inputs' selection to predict the output of interest. The algorithm can be time-consuming, depending on the number of iterations and cycles defined by the user. To compare to other renowned methodologies, models will be fitted using *Ridge,* and *LassoLars* Regression functions from the SciKit Learn linear model library. Both functions use all available inputs in one step to fit their models and are usually instantly solved, without stochastic iterations as AnTSbe. The *Ridge* Regression uses *l*2 regularization to minimize parameters, but not to absolute zero (this way, variables are not selected, only have their influence minimized). *LassoLars* Regression uses *l*1 regularization, which can drive parameters to zero, and, so, will cut variables from the final model. The AnTSbe implementation used in this work will also use *LassoLars* internally in the optimization routine. As the parsimonious criterion (*i.e.,* having a model with the least number of parameters that can still be reliable and accurate) is essential, especially when considering the expansion of the original inputs, the number of non-zero parameters in each model will be considered meaningful when comparing the fitting techniques.

After the fitting of the AnTSbe, *Ridge,* and *LassoLars* models using the calibration subset, the models are used to predict the outputs of interest with the test subset and evaluated with standard metrics: coefficient of determination ($R^2$), root mean squared error (RMSE), mean absolute percentage error (MAPE) and maximum absolute percentage error (Max e%) (GREENE, 2002), defined as follows

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \qquad RMSE = \sqrt{\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2}$$

$$MAPE = \frac{100\%}{n}\sum_i^n \left|\frac{(y_i - \hat{y}_i)}{y_i}\right| \qquad Max\ e\% = Max\left(100\% \cdot \left|\frac{y_i - \hat{y}_i}{y_i}\right|\right)$$

where $y$ is real output, $\hat{y}$ the predicted output and $\bar{y}$ the mean value of $y$.

## 6.3 Results and Discussions

### *6.3.1 Column T-01*

For Column T-01, the main variable to be predicted is the concentration of heavies (C4+) in the distillate stream, as it is the impurity that will be carried to Column T-02. This

impurity must be kept as low as possible, considering the thermodynamical and economical characteristics of the Unit. The mean, standard deviation, minimum, maximum and first, second and third quartiles of the mass ratio of heavies in the distillate stream ( $Z_{C4+}^{D1}$) can be seen in Table 6.2.

Table 6.2. Description of the main output of interest in T-01, the mass ratio of heavies in the distillate stream ( $Z_{C4+}^{D1}$).

| Output variable | Mean | Standard deviation | Minimum | First quartile | Second quartile | Third quartile | Maximum |
|---|---|---|---|---|---|---|---|
| $Z_{C4+}^{D1}$ | 0.1067 | 0.1555 | 0.00049 | 0.00071 | 0.00087 | 0.170 | 0.5255 |

The available inputs variables for Column T-01 are depicted in Table 6.3.

Table 6.3. Description of the 9 available model inputs for Column T-01.

| Variable | Description |
|---|---|
| $D/F_1$ | Mass ratio between the distillate stream flow rate and the input flow rate |
| $RR_1$ | T-01 reflux ratio |
| $Qref_1$ | Heat exchanged in the reboiler of T-01 |
| $Qcond_1$ | Heat exchanged in the condenser of T-01 |
| $F_{D1}$ | Distillate stream mass flow of T-01 |
| $F_{B1}$ | Bottom stream mass flow of T-01 |
| $T_{D1}$ | Temperature in the top of T-01 |
| $T_{B1}$ | Temperature in the bottom of T-01 |
| $DP_1$ | Pressure difference of T-01 |

The use of the Hotelling $T^2$ statistics did not identify any *outliers* in the data. The data was split into calibration and test subsets using *k-rank* with 2 centroids, in a proportion of 2:1.

The first preliminary study developed was to use an exhaustive search algorithm to evaluate if the original 9 inputs variables were able to predict the output satisfactorily $Z_{C4+}^{D1}$. Although presenting a $R^2$ around 0.99, the MAPE of the best-fitted model was 882%, and the Max e% superior to 8900%. Samples with very low concentration highly accentuated the proportional errors, and to measure impurities in T-01 distillate stream, these results were considered not ideal. Following the proposed methodology, the second and third-order polynomial expansion of the inputs will be applied for all subsequent studies. For T-01, the expansion created 210 new inputs, adding up to a total of 219 variables. Considering the size of the new dataset, the use of exhausted search is no longer viable, and the AnTSbe, *Ridge,* and *LassoLars* will be applied for variable selection and model fitting.

Table 6.4 presents the predictive metrics for the test subset of the fitted models using the three methods. It is shown that, although still high in percentage errors, the use of expanded variables enhanced the predictive capability of the models. Comparing the AnTSbe non-zero coefficients with *Ridge*, we can see a considerable reduction in model parameters with similar metrics. In this case, the AnTSbe trained models using between 2 and 30 variables, selecting the best one (24) as the one where an increase in input variables (25+) did not significantly improve the predictive metrics.

Table 6.4. Comparative predictive metrics of the whole test subset for the output $Z_{C4+}^{D1}$ in Column T-01, using expanded input variables.

| Method | Non-zero parameters | $R^2$ | RMSE | MAPE | Max e% |
|---|---|---|---|---|---|
| *LassoLars* | 59 | 0.999 | 0.0015 | 79% | 304.1% |
| *Ridge* | 219 | 0.999 | 0.0012 | 72.3% | 329.5% |
| AnTSbe | 24 | 0.999 | 0.0019 | 78% | 482% |

Further exploring the percentage errors. we could ascertain that there was a significant adhesion difference between the models and the actual outputs for samples with $Z_{C4+}^{D1}$ mass ratios higher than 0.01 and lower than 0.01. This difference can be seen in Figure 6.2, where *Ridge* model adhesion to the real outputs is presented.

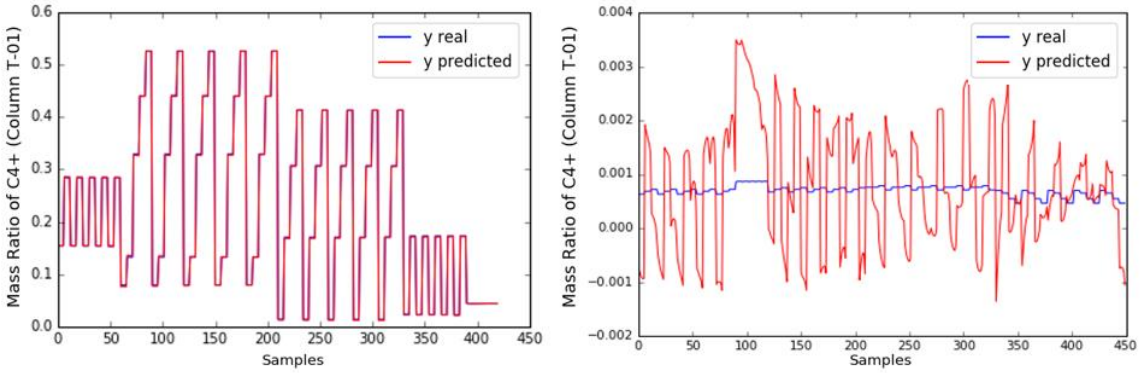



Figure 6.2. Adhesion of the *Ridge* model to the real outputs $Z_{C4+}^{D1}$, divided by samples with mass ratios above 0.01 (left) and below 0.01 (right).

Above 0.01 kg/kg the MAPE was inferior to 2% to all methods, with Max e% smaller than 5%, but the same was not true for lower concentrations. Considering the importance of rightfully asserting low concentrations for the impurities in T-01 distillate stream, local models were developed focused only on samples with $Z_{C4+}^{D1}$ smaller than 0.01. The comparative predictive metrics can be seen in Table 6.5.

Table 6.5. Comparative predictive metrics of the test subset for the output $Z_{C4+}^{D1}$ ,in concentrations lower than 0.01 kg/kg, using expanded input variables.

| Method | Non-zero parameters | $R^2$ | RMSE | MAPE | Max e% |
|---|---|---|---|---|---|
| *LassoLars* | 14 | 0.92 | 2.9e-5 | 3.3% | 9.32% |
| *Ridge* | 219 | 0.999 | 2e-6 | 0.23% | 0.96% |
| AnTSbe | 11 | 0.951 | 2.1e-5 | 2.14% | 7.40% |

The use of local models, specific for concentrations below 0.01 kg/kg, greatly enhanced their ability to predict the output correctly $Z_{C4+}^{D1}$. Both the adhesion of the *Ridge* and the AnTSbe local models can be seen in Figure 6.3.

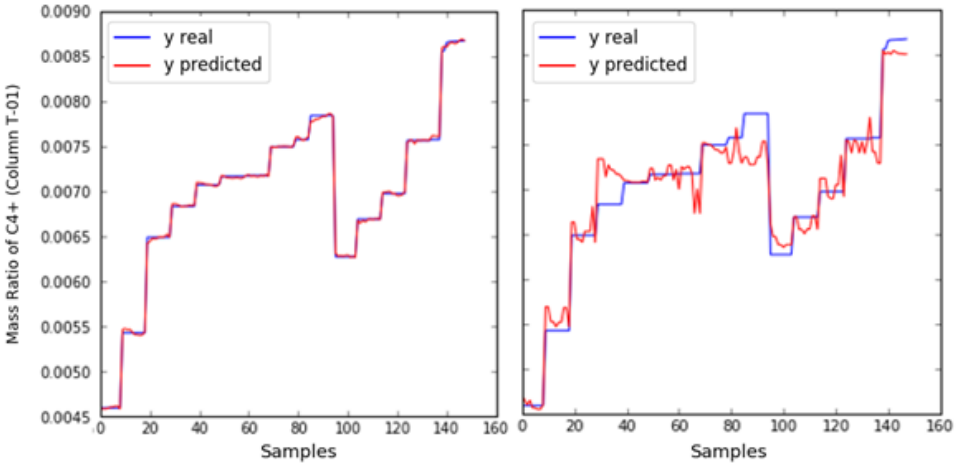



Figure 6.3. Adhesion of the *Ridge* (left) and the AnTSbe (right) local models to the real $Z_{C4+}^{D1}$ of Column T-01, for samples with heavy hydrocarbons concentration bellow 0.01 kg/kg.

Considering the range of the data and the adhesion of both the global models as the local models, we can recognize that the proposed methodology was satisfactory in predicting the main output of interest in Column T-01. The global models can also be used as region classifiers, *i.e.*, if the predicted values fell in concentrations above 0.01 kg/kg (or a user-defined threshold), they are considered trustworthy. Otherwise, if the global model-predicted concentrations below 0.01 kg/kg, the local models are used to replace the results for more accurate predictions.

As a final study of Column T-01, the concentration of the other constituents of the distillate stream (ethane – $Z_{C2}^{D1}$, propylene – $Z_{C3-}^{D1}$ and propane – $Z_{C3+}^{D1}$) were also predicted using the AnTSbe method. As this stream directly feeds Column T-02, this information can be helpful for Unit operation. Table 6.6 presents the predictive test metrics for these outputs. All concentrations could be satisfactorily predicted in the distillate stream, been suited to be used as inputs for Column T-02. Considering all adjusted models, the concentration profile of the distillate stream can be followed and may even be the basis for data reconciliation strategies.

Table 6.6. Predictive metrics for other important constituents of the distillate stream in Column T-01 (AnTSbe method).

| Output | Non-zero parameters | $R^2$ | RMSE | MAPE | Max e% |
|---|---|---|---|---|---|
| Ethane – $Z_{C2}^{D1}$ | 27 | 0.999 | 0.00058 | 0.96% | 3.18% |
| Propylene – $Z_{C3-}^{D1}$ | 23 | 0.999 | 0.0041 | 0.48% | 1.57% |
| Propane – $Z_{C3+}^{D1}$ | 16 | 0.999 | 0.0016 | 0.57% | 1.90% |

### *6.3.2 Column T-02*

For Column T-02, it is economically essential to reduce the waste of propylene in the distillate stream. The mean, standard deviation, minimum, maximum, and first, second and third quartiles of the mass ratio propylene in the distillate stream ( $Z_{C3-}^{D2}$) can be seen in Table 6.7.

Table 6.7. Description of the main output of interest in T-02, the mass ratio of propylene in the distillate stream ( $Z_{C3-}^{D2}$).

| Output variable | Mean | Standard deviation | Minimum | First quartile | Second quartile | Third quartile | Maximum |
|---|---|---|---|---|---|---|---|
| $Z_{C3-}^{D2}$ | 0.6176 | 0.1702 | 0.0128 | 0.5209 | 0.6516 | 0.7445 | 0.9396 |

The available inputs variables for Column T-02 are depicted in Table 6.8.

Table 6.8. Description of the 12 available model inputs for Column T-02.

| Variable | Description |
|---|---|
| $B/F_2$ | Mass ratio between the bottom stream and the input stream |
| RR2 | T-02 reflux ratio |
| $Qref_2$ | Heat exchanged in the reboiler of T-02 |
| $Qcond_2$ | Heat exchanged in the condenser of T-02 |
| $F_{D2}$ | Distillate stream mass flow of T-02 |
| $F_{B2}$ | Bottom stream mass flow of T-02 |
| $T_{B2}$ | Temperature in the bottom stream of T-02 |
| $P_{B2}$ | Pressure in the bottom stream of T-02 |
| $DP_2$ | Pressure difference of T-02 |
| $Z_{C3-}^{I2}$ | Propylene concentration in the feed of T-02 |
| $Z_{C3+}^{I2}$ | Propane concentration in the feed of T-02 |
| $Z_{C2}^{I2}$ | Ethane concentration in the feed of T-02 |

Using Hotelling $T^2$ with α equal to 1%, almost 22.000 samples were removed as outliers. Column T-02 had an abundance of data, still holding 392817 samples after outlier removal. The *k-rank* with 2 centroids was recursively applied to split the data into calibration and test subsets with different proportions, trying to find a calibration subset sufficient to represent the data without losing the model's generalization performance. It was necessary to use only 0.5% of the data as calibration (1965 samples) to represent the data faithfully. Enlarging the volume of calibration samples did not enhance the model's predictive metrics in the test subset. Before fitting the models, the 12 original inputs were expanded to 454 variables.

Table 6.9 presents the test subset comparative predictive metrics for calculating the wasted propylene in the distillate stream.

Table 6.9. Comparative predictive metrics of the test subset for the output $Z_{C3-}^{D2}$ in Column T-02, using expanded input variables.

| Method | Non-zero parameters | $R^2$ | RMSE | MAPE | Max e% |
|---|---|---|---|---|---|
| *LassoLars* | 105 | 0.996 | 0.0097 | 1.05% | 40.6% |
| *Ridge* | 454 | 0.997 | 0.0087 | 0.95% | 33.53% |
| AnTSbe | 32 | 0.991 | 0.0162 | 2.11% | 45.6% |

Figure 6.4 presents the real versus the predicted output plot for both the Ridge and the AnTSbe model. There is deviance in the prediction performance in higher concentrations. As the number of test samples is massive, and most of them had excellent performance, the MAPE is considerably low, but the Max e% shows relevant error in some samples. AnTSbe was, in general, better at distributing the error across the concentration range. As the concentration of propylene should be kept low in the distillate stream, the performance of both models can be considered satisfactory, as most of the error is focused in areas with higher concentration.

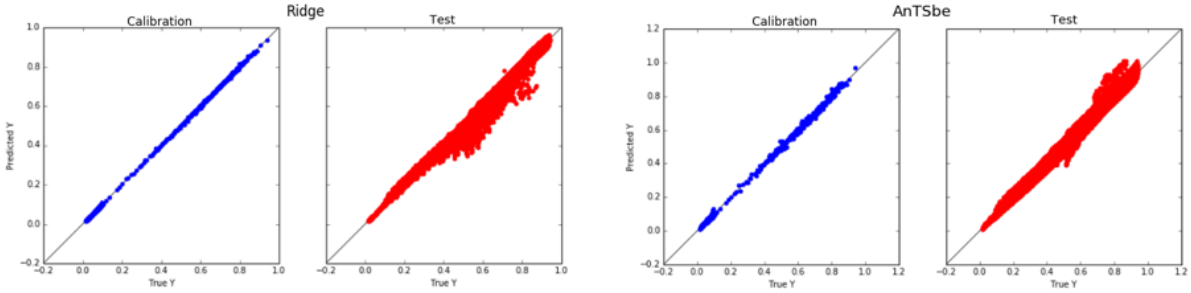



Figure 6.4. Real *versus* the predicted output of $Z_{C3-}^{D2}$ for Column T-02, for the *Ridge* model (left) and the AnTSbe model (right).

The concentration of propylene in the bottom stream of T-02 is also relevant, as it directly feeds T-03. AnTSbe was used to predict this concentration. Keeping only 27 non-zero parameters, the model could competently predict the output in the test subset with MAPE of 0.18% and a Max e% of 2.63%.

### *6.3.3 Column T-03*

In Column T-03 it is essential to maintain the specification for propylene in the distillate stream. The Unit requires that the contamination of propane in the distillate stream must be kept below 0.4% (0.004 kg/kg), holding a propylene purity of 99.6%. As impurities are harder to predict, the main output of T-03 is the mass ratio of propane in the distillate $Z_{C3+}^{D3}$. The mean, standard deviation, minimum, maximum, and first, second and third quartiles of the mass ratio propane in the distillate stream ( $Z_{C3+}^{D3}$) can be seen in Table 6.10.

Table 6.10. Description of the main output of interest in T-03, the mass ratio of propane in the distillate stream ( $Z_{C3+}^{D3}$).

| Output variable | Mean | Standard deviation | Minimum | First quartile | Second quartile | Third quartile | Maximum |
|---|---|---|---|---|---|---|---|
| $Z_{C3+}^{D3}$ | 0.1094 | 0.1064 | 1e-6 | 0.0029 | 0.0692 | 0.1953 | 0.3292 |

The available inputs variables for Column T-03 are depicted in Table 6.11. Some variables will be referred to the same nomenclature presented in Figure 6.1.

Table 6.11. Description of the 18 available model inputs for Column T-03.

| Variable | Description |
|---|---|
| $Z_{C3-}^{I3}$ | Propylene concentration in the feed of T-03 |
| $FR_3$ | Ratio of the flow that leaves the reboiler and return as reflux to the column |
| $FA_3$ | Ratio of the flow that leaves the compressor and will be used as heating fluid for the reboiler |
| $F_{B3}$ | Bottom stream mass flow of T-03 |
| $P_{B3}$ | Bottom stream pressure of T-03 |
| $F_{D3}$ | Distillate stream mass flow of T-03 |
| $T_{D3}$ | Temperature in the distillate stream of T-03 |
| $F_{RR}$ | Mass flow of the reboiler that returns to the column as reflux |
| $F_{CR}$ | Mass flow of the condenser that returns to the column as reflux |
| $W_{C\text{-}01}$ | Energy required for compressor (C-01) |
| $W_{B\text{-}04}$ | Energy required for pump (B-04) |
| $Q_{P06}$ | Energy required for condenser (P-06) |
| $Q_{P07}$ | Energy required for exchanger (P-07) |
| $T_D$ | Temperature at the top of Column T-03 |
| $T_B$ | Temperature at the bottom of Column T-03 |
| $P_B$ | Pressure at the bottom of Column T-03 |
| $V_{REF}$ | Vaporized fraction that leaves the reboiler |
| $DP_3$ | Pressure difference of T-03 |

Appling Hotelling's $T^2$, 85 samples were removed as outliers (around 5% of the data), leaving 1679 operational points. The 18 original inputs were expanded to 1330 variables, and the data was split into calibration and test subsets using *k-rank* with 2 centroids and a proportion of 2:1.

It was necessary to adopt a different strategy for Column T-03. After a preliminary study, it was impossible to predict the distillate impurity satisfactorily using only the 18 input variables and their expansions. To achieve suitable results, it was necessary first to predict the mass ratio of propylene in the distillate stream $Z_{C3-}^{D3}$, and use this information, along with the others inputs, to predict the propane contamination. The AnTSbe method was applied to predict the propylene concentration. Keeping only 6 non-zero coefficients accomplished relevant results with a $R^2$ of 0.99, RMSE of 0.019, MAPE of 0.17%, and Max e% of 0.80%, for the test subset. Figure 6.5 presents the real *versus* the predicted $Z_{C3-}^{D3}$ for the distillate stream of Column T-03.

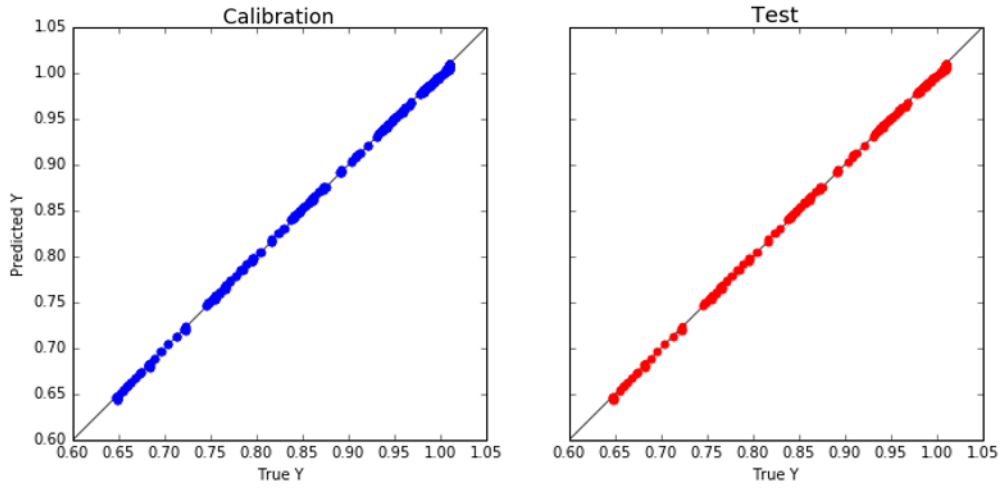



Figure 6.5. Real *versus* the AnTSbe predicted $Z_{C3-}^{D3}$ for the distillate stream of Column T-03

Confident in the predictions of propylene, this information is added to the original 18 inputs, and they are expanded to 1539 variables that will be used to forecast the concentration of the propane impurity in the distillate stream. Here, as in Column T-01, it is valid to fit local models focused on a narrower concentration range. The samples will be divided into two groups, one containing samples with more than 0.01 kg/kg of propane and the other with samples below this threshold.

For samples with $Z_{C3+}^{D3}$ higher than 0.01, all three methods presented good results, with *Ridge* and AnTSbe with MAPE in the order of 0.2% and Max e% of 3%, for the test subset. *LassoLars* presented errors one order of magnitude higher. Detailed results of this sample group will be suppressed as they are not relevant for the Unit's operation. As the Unit specification requires no more than 0.004 kg/kg of propane in the distillate, any prediction greater than 0.01 kg/kg would automatically be red-flagged, and the product would need to be reworked.

Table 6.12 presents the comparative performance of the models in predicting the propane impurity in the distillate stream of Column T-03, in samples with $Z_{C3+}^{D3}$ lower than 0.01, for the test subset.

Table 6.12. Comparative metrics for the prediction of the propane impurity in the distillate stream of Column T-03, $Z_{C3+}^{D3}$ ,for propane concentrations lower than 0.01 kg/kg, for the test subset.

| Method | Non-zero parameters | $R^2$ | RMSE | MAPE | Max e% |
|---|---|---|---|---|---|
| *LassoLars* | 117 | 0,949 | 0.0003 | 530% | 10298% |
| *Ridge* | 1539 | 0,996 | 8.5e-5 | 180% | 6336% |
| AnTSbe | 17 | 0,999 | 2.1e-5 | 83% | 2910% |

Although low in RMSE and high in $R^2$, there is a clear disbalance in the percentage errors, even using local models. A meticulous analysis shows that the percentage errors grow excessively as the propane concentration approaches zero. Figure 6.6 shows the absolute percentage errors for two separate regions of concentrations: $Z_{C3+}^{D3}$ between 0.0002 to 0.01 and for samples below 0.0002, using the AnTSbe fitted model.

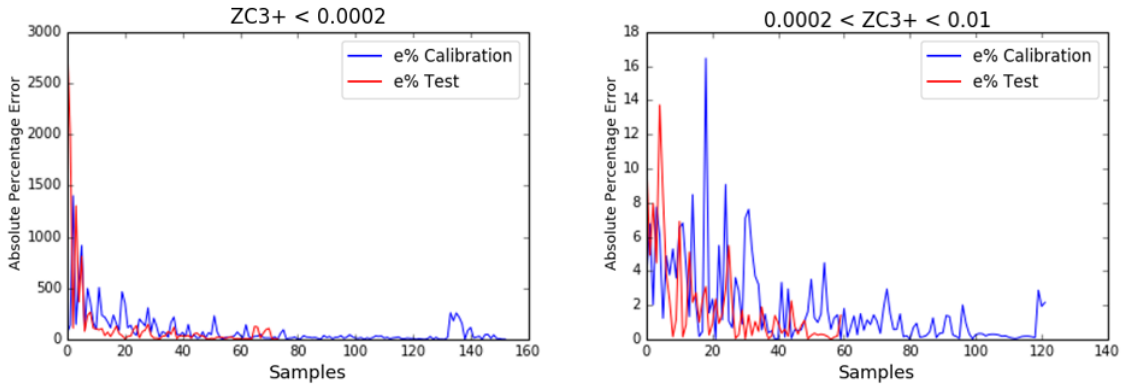



Figure 6.6. Individual percentage errors for T-03 AnTSbe fitted model, for samples with $Z_{C3+}^{D3}$ lower than 0.0002 (left) and between 0.0002 to 0.01 (right).

Evaluating the plots in Figure 6.6, shows that most of the percentage error happens in the region where impurity concentration is below 0.0002 kg/kg. In the range between 0.0002 to 0.01 kg/kg, the Max e% is 16%, with a MAPE smaller than 5%. This result is relevant once, considering the Unit's specification, this is the region of greater interest for the operators. The adhesion of the AnTSbe model, in this concentration range, to the original outputs can be seen in Figure 6.7.

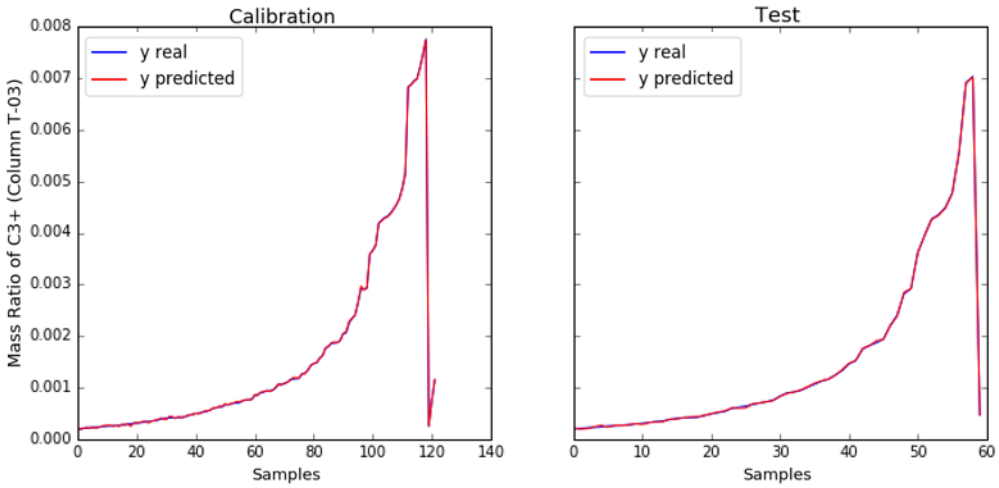



Figure 6.7. Column T-03 AnTSbe model's adhesion to the real outputs $Z_{C3+}^{D3}$, for samples with propane concentration between 0.0002 and 0.01 kg/kg

Figure 6.8 presents the AnTSbe model's adhesion to the real outputs for samples with propane concentration below 0.0002 kg/kg. Even considering the maximum absolute percentage error of almost 3000%, the plots show that this error is not detrimental for the Unit's operation, once, even in this worst-case scenario, the distillate stream would be far from reaching the maximum concentration of propane accepted in the final product.

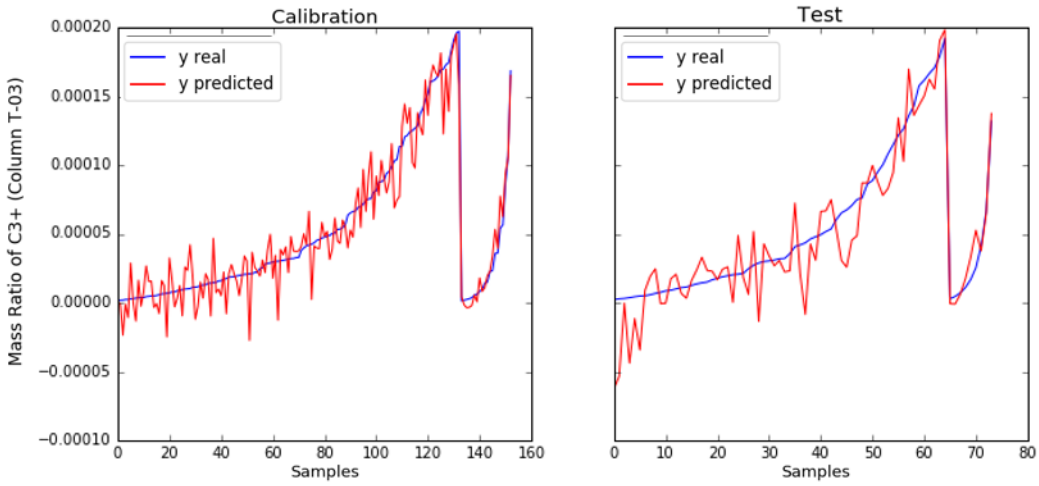



Figure 6.8. Column T-03 AnTSbe model's adhesion to the real outputs $Z_{C3+}^{D3}$, for samples with propane concentration below 0.0002 kg/kg

## 6.4 Conclusions

Petroleum refining has been one of the key chemical engineering processes for more than a century, reaching extremely high product specification and process control standards. Although extensively studied, crucial process variables are still tricky, costly, or time-consuming to measure daily, primarily related to product quality. Propylene is a common subproduct of petroleum cracking that has gained significant importance, and on-purpose production processes have been on the rise. There is an international gap between supply and demand, and the recovery of high purity propylene from liquefied petroleum gas can be profitable. Irrespective of the propylene manufacturing route, the product is always a mixture of propylene/propane that must be separated. In this work, we proposed a methodology to develop black-box linear soft-sensors for a propylene/propane splitter unit, capable of predicting key process variables from the polynomial expansion of readily available information.

For Columns T-01, the main variable to be predicted is the concentration of heavies (C4+) in the distillate stream, as it is the impurity that will be carried to Column T-02. As the concentration must be kept low, it was necessary to create local models, dividing the operational points into samples with $Z_{C4+}^{D1}$ mass ratios higher and lower than 0.01. Above 0.01 kg/kg the MAPE was inferior to 2% to all methods, with Max e% smaller than 5%. For samples with $Z_{C4+}^{D1}$ smaller than 0.01 (more relevant to the operation of the Unit), the AnTSbe model kept 11 out of the 219 available inputs, with RMSE of 2.1e-5, MAPE of 2.14% and Mas e% of 7.40%. It was also possible to predict the concentrations of ethane – $Z_{C2}^{D1}$, propylene – $Z_{C3-}^{D1}$ and propane – $Z_{C3+}^{D1}$ into the distillate stream satisfactorily, with Max e% smaller than 3%. This information is essential because this is the stream that feeds Column T-02.

For Column T-02, it is economically essential to reduce the waste of propylene in the distillate stream. The AnTSbe model kept 32 out of the 454 variables and predicted the wasted propylene concentration in the distillate with RMSE of 0.0162, MAPE of 2.11%, but a high Max e% of 45.6%. Evaluating the distribution of the errors along with the range of concentrations, it was clear that the models have been less precise into areas of higher concentration of propylene, and, as this concentration must be kept low in the operation

of the Unit, the model performance can be considered sufficient, as it presented minor errors for lower concentrations.

In Column T-03 it is essential to maintain the specification of 99.6% purity of propylene in the distillate stream. The propane contamination in the distillate stream must be kept below 0.4% (0.004 kg/kg). The main output of T-03 is the mass ratio of propane in the distillate $Z_{C3+}^{D3}$. A different approach was needed for the column, first using the 18 original variables to predict the mass ratio of propylene in the distillate stream ($Z_{C3-}^{D3}$) and then adding this variable to the inputs to predict the propane impurity. The AnTSbe model using 6 out of the 1330 available expanded inputs accomplished relevant results in predicting the mass propylene ratio with a $R^2$ of 0.99, RMSE of 0.019, MAPE of 0.17%, and Max e% of 0.80%, for the test subset. Once again, local models were applied for $Z_{C3+}^{D3}$ higher and lower than 0.01. For concentrations of propane lower than 0.01 kg/kg, the AnTSbe model kept 17 out of the 1539 available variables and predicted the impurity of interest with RMSE of 2.1e-5, MAPE of 83%, and Max e% of 2910%. A meticulous evaluation of the individual errors showed that the great majority of the percentage errors were focused in the region with $Z_{C3+}^{D3}$ lower than 0.0002. For the range of mass ratio between 0.0002 and 0.01 (region of most significant interest for the Unit's operation), the MAPE error is lower than 5%, and the Max e% is 16%. Even in the worst-case predictive scenario, for samples with propane concentration lower than 0.0002 kg/kg, and maximum errors of 3000%, the distillate stream would be far from reaching the maximum concentration of propane accepted in the final product. In Column T-03, with 1539 available inputs, we can see a clear advantage of the AnTSbe method against the *Ridge* and *LassoLars* direct application.

## 6.5 References


AKAH, A.; AL-GHRAMI, M. Maximizing propylene production via FCC technology. **Applied Petrochemical Research**, v. 5, n. 4, p. 377–392, 22 dez. 2015.

BROOKS, R. **Modeling the North American Market for Natural Gas Liquids**. 2013

BRYAN, P. F. Removal of Propylene from Fuel-Grade Propane. **Separation & Purification Reviews**, v. 33, n. 2, p. 157–182, 12 jan. 2004.

DIMIAN, A. C.; BILDEA, C. S.; KISS, A. A. Methanol-To-Olefin Process. In: **Applications in Design and Simulation of Sustainable Chemical Processes**. [s.l.] Elsevier, 2019. p. 147–182.

ECHEMI. **PROPYLENE Price Market Analysis**.

GLOVER, F. Tabu Search—Part I. **ORSA Journal on Computing**, v. 1, n. 3, p. 190–206, 1 ago. 1989.

GPSA. **Gas Processors Suppliers Association**. 12. ed. Tulsa, OK: [s.n.].

GRAZIANI, S.; PAGANO, F.; XIBILIA, M. G. **Soft sensor for a Propylene Splitter with seasonal variations**. 2010 IEEE Instrumentation Measurement Technology Conference Proceedings. **Anais**...2010

GREENE, W. H. **Econometric analysis**. Upper Saddle River, N.J.: Pearson Education, 2002.

HEINRITZ-ADRIAN, M.; WENZEL, S.; YOUSSEF, F. Advanced propane dehydrogenation. **Petroleum Technology Quarterly**, v. 13, p. 83,85-86,88,91, 1 jan. 2008.

HINOJOSA, A. I.; CAPRON, B.; ODLOAK, D. REALIGNED MODEL PREDICTIVE CONTROL OF A PROPYLENE DISTILLATION COLUMN. **Brazilian Journal of Chemical Engineering**, v. 33, p. 191–202, 2016.

HOTELLING, H. **Analysis of a complex of statistical variables into principal components**.

Baltimore: Warwick & York, 1933.

LAVRENOV, A. V. et al. Propylene production technology: Today and tomorrow. **Catalysis in Industry**, v. 7, n. 3, p. 175–187, 29 jul. 2015.

MAUHAR; BARJAKTAROVIC; SOVILJ, M. Optimization of Propylene-Propane Distillation Process. **Chemical Papers**, v. Vol. 58, N, 1 jan. 2004.

PALMER, E. et al. High purity propylene from refinery LPG. v. 17, 1 jan. 2012.

PAQUET-DURAND, O. et al. Artificial neural network for bioprocess monitoring based on fluorescence measurements: Training without offline measurements. **Engineering in Life Sciences**, v. 17, n. 8, p. 874–880, 9 out. 2017.

PĂTRĂŞCIOIU, C.; ANH, C.; POPESCU, M. **Control of propylene - propane distillation process using Unisim® design**. [s.l: s.n.].

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, n. Oct, p. 2825–2830, 2011.

RANZAN, L.; TRIERWEILER, L. F.; TRIERWEILER, J. O. Prediction of sulfur content in diesel fuel using fluorescence spectroscopy and a hybrid ant colony - Tabu Search algorithm with polynomial bases expansion. **Chemometrics and Intelligent Laboratory Systems**, v. 206, p. 104161, 15 nov. 2020.

SCHULTZ, E. S. **A importância do ponto de operação nas técnicas de Self-optimizing Control** (Ufrgs, Ed.)Porto Alegre, RS, Brazil, 2015.

STÜTZLE, T.; LÓPEZ-IBÁÑEZ, M.; DORIGO, M. A Concise Overview of Applications of Ant Colony Optimization. In: [s.l: s.n.].

TIBSHIRANI, R. **Regression Shrinkage and Selection Via the Lasso**. Journal of the Royal Statistical Society, Series B. **Anais**...1994

UMO, A. M.; BASSEY, E. N. Simulation and Performance Analysis of Propylene-Propane Splitter in Petroleum Refinery Case Study. **International Journal of Chemical Engineering and Applications**, v. 8, n. 1, p. 1–4, fev. 2017.

VENKATASUBRAMANIAN, V. et al. A review of process fault detection and diagnosis part I: Quantitative model-based methods. **Computers and Chemical Engineering**, v. 27, n. 3, p. 293–311, 15 mar. 2003.

VENKATESH BABU, S.; RAMESH, G. Recovery of Propylene from LPG. **International Journal of Latest Trends in Engineering and Technology**, v. Volume 11, n. Issue 2, 2013.

XU, L.; ZHANG, W. J. **Comparison of different methods for variable selection**. Analytica Chimica Acta. **Anais**...Elsevier, 19 nov. 20

# Capítulo 7 – **Conclusões e Trabalhos Futuros**

## 7.1 Conclusões

O estudo de modelos baseados em dados provenientes de espectroscopia por fluorescência se mostrou de grande valia para diversas aplicações no universo industrial. Apesar de amplamente pesquisados, ainda se identifica baixa penetrabilidade de sensores óticos em operações de larga escala nos ramos adjacentes da engenharia química. Um dos possíveis motivos se deve a complexidade no tratamento destes dados, que requer uma gama de técnicas especializadas para transformar a informação contida nos espectros em conhecimento útil de processo.

Neste âmbito, as pesquisas desenvolvidas neste trabalho são continuidade natural do desenvolvimento de sensores de processo personalizados, baseados em técnicas óticas. As vantagens associadas a este tipo de sensor são especialmente relevantes para processos químicos e biológicos. A coleta espectral é realizada de maneira rápida, não intrusiva, não destrutiva e não requer tratamento da amostra. Além disso, são técnicas sensíveis e com capacidade de correlacionar um único espectro a uma miríade de propriedades do meio. Em especial para espectroscopia por fluorescência, uma quantidade considerável dos analitos de interesse nestas áreas são fluoróforos naturais.

A metodologia para seleção de variáveis explicativas proposta neste trabalho, AnTSbe, se mostrou útil na área de quimiometría, resultando em modelos com capacidade preditiva superiores a metodologias consolidadas, como PLS, e também a versões anteriores de algoritmos Colônia de Formigas, quando comparados utilizando a mesma base de dados (quantificação de enxofre total em amostras de diesel combustível). Além de apresentar as variáveis selecionadas para o modelo com menores erros de predição, o algoritmo expõe, através das trilhas de feromônios, um ordenamento das variáveis de entrada com maior correlação com as saídas de interesse. Essa informação é proveitosa para técnicas de filtragem de dados, especialmente vantajosas para situações onde a base de dados é extensa e altamente colinear.

A introdução do conceito de lista tabu se mostrou benéfico para metodologia, evitando, no melhor dos casos, uma repetição redundante de 20% de todos os subgrupos testados pelo algoritmo. A real influência da hibridização com Busca Tabu é ainda maior do que a simples contagem de vezes que o algoritmo foi impedido de selecionar um grupo proibido: sem esta restrição, a ciclagem de subgrupos elevaria a quantidade de feromônios nestas variáveis com alta probabilidade de seleção, causando uma possível estagnação precoce. A introdução de subgrupos na lista de proibição garante, também, que a evaporação da trilha ao final de cada iteração melhor regule a relação de feromônios entre as variáveis de entrada.

A natureza dos dados de fluorescência acarreta em um detrimento nos benefícios da hibridização, uma vez que pares de fluorescência adjacentes carregam, muitas vezes, informação similar. Quanto menor o número de variáveis de entrada, e mais independentes elas forem, maior é a vantagem no uso da lista tabu, evitando desperdício de tempo computacional e melhorando os mecanismos de prospecção e exploração, tornando a metodologia mais consistente e eficiente.

A introdução da expansão polinomial e combinatória das variáveis de entrada, no interior da rotina de otimização, tornou possível a captura de uma maior gama de informação dos dados, melhorando as métricas de predição para modelos que não apresentavam bons resultados apenas com a base original. Este efeito pôde ser observado no estudo de caso dos modelos locais para Diesel S100, e é fortemente evidenciado para as predições de compostos fenólicos em cachaça envelhecida, assim como para inferidores de estado nas correntes da unidade de separação de propano/propeno. A introdução da expansão acarretou em uma diminuição de até 50% dos erros médios de predição de fenólicos totais, quando não consideradas amostras comerciais de cachaça. Para a predição dos perfis de concentração e impurezas da USPP, o uso da expansão de bases foi essencial para construção dos modelos, sendo inclusive desconsiderados os modelos sem seu uso.

O uso da expansão de bases, apesar de benéfico para a maioria dos modelos ajustados, evidenciou uma fragilidade na técnica. A expansão da ordem dos dados também ocasiona um aumento da sensibilidade dos modelos à distúrbios não medidos durante a etapa de calibração, fazendo com que os mesmos possuam menor poder de generalização. Isto pode ser observado quando amostras de cachaça comercial, bem distintas das amostras envelhecidas em laboratório, participavam dos subgrupos de teste para os modelos preditivos de fenólicos totais, com erros muito superiores as demais amostras.

O desenvolvimento da metodologia Adaptada AnTSbe, com introdução do conceito de Par Delta, obteve sucesso em aperfeiçoar a performance preditiva de modelos que utilizam expansão de base. O uso de um par de fluorescência como regulador do meio aprimorou a capacidade dos modelos ajustados em predizer amostras que apresentavam distinções significativas do grupo de calibração. Na predição de compostos fenólicos totais em amostras de cachaça envelhecida, o uso da metodologia adaptada apresentou erros até três vezes menores, em comparação a metodologia original, quando considerado o subgrupo de teste contendo cachaças comerciais. Excluindo as amostras comerciais, ambas metodologias apresentaram resultados satisfatórios similares nos subgrupos de teste.

A caracterização dos perfis de compostos fenólicos em processos de envelhecimento de cachaças em barris de madeira propiciou múltiplas contribuições. Primeiramente, não é corriqueiro encontrar na literatura trabalhos que acompanham processos de envelhecimento com a coleta frequente de espectros por fluorescência como o

apresentado. Isso se deve principalmente a dinâmica temporal necessária, tendo a pesquisa percorrido mais de três anos para sua conclusão. Este acompanhamento identificou as radicais mudanças nos espectros que acontecem ao longo do envelhecimento, como o deslocamento evidente dos picos de fluorescência. Outro fato interessante a ser pontuado foram as diferenças nas intensidades médias de fluorescência dos espectros entre os diferentes envelhecimentos.

Tanto os modelos ajustados com a metodologia AnTSbe original, quanto com a adaptada, foram capazes de predizer satisfatoriamente a concentração de fenóis ao longo dos envelhecimentos. A análise dos dados e dos resultados mostrou necessário o ajuste de modelos locais que separassem os envelhecimentos CA2 e CA3 das amostras CA1, uma vez que tanto os espectros quanto as concentrações dos dois grupos eram significativamente distintas. A metodologia adaptada, por sua vez, foi mais eficiente em predizer amostras comerciais de cachaça envelhecida, onde o algoritmo original foi falho. Ambas abordagens se mostraram úteis como base para o desenvolvimento de sensores capazes de refinar processos de envelhecimento, com acompanhamento do perfil fenólico, do tempo de vida útil dos barris, e um maior controle sobre a padronização do produto final.

Com relação a caracterização de bioprocessos, o uso de redes neurais convolucionais residuais se mostrou eficiente em representar de forma fiel as concentrações de interesse no meio ao longo dos processos fermentativos. Para as três fermentações estudadas, os erros de predição foram mínimos, com $R^2$ superiores a 0.99. O uso de redes convolucionais com dados bidimensionais de fluorescência apresentou alto poder preditivo, e tem potencial para ser aplicado em outras pesquisas quimiométricas. Vale ressaltar que a qualidade do modelo neural está intimamente ligada a abundância de dados que o acoplamento de um espectrofluorômetro no sistema fermentativo possibilitou.

A metodologia de triagem e detecção de inconformidades baseada em redes autoencoder também se mostrou válida. A rede calibrada foi capaz de reconstruir os espectros de entrada dos sistemas fermentativos com erros médios inferiores a 0.65%. Utilizando a rede calibrada, e acompanhando o erro de reconstrução de um novo conjunto de pontos modificados artificialmente, foi possível identificar as amostras que receberam erros significativos, e também quantificar quanto essas amostras divergiam dos dados de calibração originais, baseado na magnitude do erro. Corroborando com a hipótese levantada, as amostras identificadas como anormais apresentaram erros médios de predição superiores aos esperados, quando preditos pela rede residual calibrada com a mesma base de dados original. O uso da metodologia pode atenuar um dos desafios no emprego de modelos empíricos, qualificando não supevisionadamente se novos dados a serem rodados por um modelo se enquadram nas características dos dados utilizados para sua calibração. Um influxo de novas leituras com grandes erros de reconstrução pode sinalizar algum problema na coleta de dados, ou então que o novo estado do sistema não é mais bem representado pelo grupo de calibração e que o modelo deve ser reajustado.

A expansão da aplicação da metodologia AnTSbe para dados industriais também apresentou bons resultados. A partir dos dados simulados baseados em uma unidade real de separação propano/propeno, foi possível predizer as concentrações dos compostos de interesse em cada uma das três colunas de separação com precisão satisfatória. Os modelos ajustados para predição das impurezas nas correntes principais de cada coluna

(fundamental interesse do trabalho) demandaram o uso de modelos locais especificados por faixas de concentração, e atingiram predições com erros absolutos médios inferiores a 5%. Para a terceira coluna, onde é necessário o controle da especificação do produto final, a metodologia AnTSbe foi capaz de lidar com a grande quantidade de dados de entrada expandidos (superiores a 1500) de maneira mais eficiente que as demais técnicas apresentadas (Regressão *Ridge* e *LassoLars*).

De maneira geral, pode-se concluir com este trabalho que as metodologias propostas apresentam boa viabilidade para aplicação na caracterização de processos industriais. A combinação de espectroscopia por fluorescência bidimensional com técnicas de seleção de variáveis e redes neurais podem ser a base para construção de novos sensores capazes de mensurar rapidamente informações que hoje são dependentes de testes laboratoriais, abrindo possibilidades para novas estratégias de controle de processos. O uso de redes autoencoder, por sua vez, pode ter sua aplicação expandida dentro do ramo quimiométrico, evitando predições falhas e apurando a confiabilidade dos sistemas preditivos.

## 7.2 Sugestões para Trabalhos Futuros

As metodologias desenvolvidas neste trabalho para caracterização de processos a partir de dados de espectroscopia por fluorescência bidimensional se mostraram satisfatórias no intuito de selecionar e transcrever a informação espectral em conhecimento útil de processo. Apesar disso, diversas são as possibilidades de evolução e melhoria das técnicas.

A metodologia AnTSbe, ainda que apresentada majoritariamente neste trabalho com dados provenientes de espectroscopia por fluorescência, pode ser utilizada para a otimização de modelos e seleção de variáveis com quaisquer tipos de dados. Além disso, as matrizes de dados de espectroscopia podem ser acrescidas de outras fontes de informação, como, por exemplo, medidas de pH, temperatura e pressão, que já estão normalmente disponíveis no sistema. O algoritmo iria avaliar, então, se existe benefício em adicionar estas variáveis nos modelos de predição. Dados de espectroscopia vibracional também podem ser incluídos, uma vez que sua aquisição e benefícios muito se assemelham a fluorescência, aumentando a gama de informações disponíveis.

Outras hibridizações também podem ser apresentadas para ampliar as funcionalidades do algoritmo. Nos casos discutidos, a seleção de variáveis é realizada de forma discreta. Uma rotina interna de otimização utilizando meta-heurísticas como Enxame de Partículas pode ser sugerida para solução de problemas contínuos, inclusive com a proposição de ajuste de modelos não lineares.

O uso de redes residuais se provou eficiente para a caracterização de bioprocessos. Porém, sua aplicação está atrelada a disponibilidade de dados. Este é mais um incentivo para aprofundar estudos que apliquem coleta *online* de dados espectrais, com o acoplamento de espectrofluorômetros diretamente no sistema a ser estudo. O influxo de uma maior quantidade dados espectrais é o caminho natural para confirmar de forma prática as funcionalidades teóricas apresentadas neste trabalho para dados laboratoriais, como para caracterização de produtos petroquímicos.

Com relação a aplicação de redes autoencoder para identificação de inconformidades, é sugerida a ampliação de estudos com o emprego em diferentes casos de estudos, para

explorar sua funcionalidade. A metodologia, em teoria, pode ser associada a modelos quimiométricos existentes de diferentes áreas, principalmente aquelas que possuem grande variabilidade de matéria-prima ou processual.

Considerando as características dos dados de fluorescência, outras abordagens para redes autoencoder podem ser aplicadas. Redes autoencoder do tipo *denoising* (geralmente construídas com arquitetura de redes convolucionais) podem ser utilizadas para, ao invés de identificar anormalidades, corrigir automaticamente espectros falhos e remover ruído da matriz de dados espectrais. Redes do tipo *variational* também podem ser uteis. Estas redes são análogas as redes autoencoder tradicionais, porém, ajustam um modelo generativo a partir da seção de decodificação, capaz de criar novos dados de entrada com características específicas. Essas redes são muito utilizadas para mesclar características dos dados. Esta abordagem é proveitosa quando a base de dados é restrita, tornando possível a criação computacional de uma pletora de espectros com ampla variabilidade, que seriam custosos ou impraticáveis laboratorialmente, a partir de uma quantidade reduzida de amostras chave.

Além das modificações conceituais sugeridas, é proposto a aplicação das metodologias para caracterização de outros processos com base em dados de espectroscopia, propiciando um melhor aproveitamento das ferramentas desenvolvidas e incentivando o progresso de técnicas óticas. Neste caso, aventamos o enfoque em sensores customizados, com coleta inteligente de dados (baseada nos princípios e vantagens da seleção de variáveis explicativas, amplamente debatidas neste trabalho). Com menor custo associado e uma maior especificidade, o uso destes sensores pode propiciar a automação de processos que atualmente não viabilizam economicamente o investimento em instrumentos ópticos especializados.